Universidade Federal de Santa Catarina

Programa de Pós-Graduação em Engenharia de Produção

MARCEL HUGO

UMA INTERFACE DE RECONHECIMENTO DE VOZ PARA O SISTEMA DE GERENCIAMENTO DE CENTRAL DE INFORMAÇÃO DE FRETES.

Dissertação submetida à Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Engenharia.

Florianópolis, Setembro de 1995.

UMA INTERFACE DE RECONHECIMENTO DE VOZ PARA O SISTEMA DE GERENCIAMENTO DE CENTRAL DE INFORMAÇÃO DE FRETES.

MARCEL HUGO

Esta dissertação foi julgada adequada para obtenção do Título de "Mestre em Engenharia".

Especialidade em Engenharia de Produção e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção.

Osmar Possamai, Dr.
Coordenador do Curso

Banca Examinadora:

Ricardo Miranda Barcia, Ph.D.
Presidente

Édis Mafra Lapolli, Dr.

Lia Caetano Bastos, Dr.

"Estudar é, realmente, um trabalho difícil. Exige de quem o faz uma postura crítica, sistemática. Exige uma disciplina intelectual que não se ganha a não ser praticando-a."

"Estudar não é um ato de consumir idéias, mas de criá-las e recriá-las."

Paulo Freire

A meus pais e Isabella.

AGRADECIMENTOS

Gostaria de agradecer o apoio e confiança depositadas em mim pelos professores Ricardo Miranda Barcia e Paulo de Tarso Mendes Luna, desde o momento em que tive a intenção de ingressar no Programa de Pós-Graduação em Engenharia de Produção até a obtenção do titulo almejado.

Obrigado ao amigo Prof. Sérgio Stringari, pelo conselhos e sugestões.

Agradecimentos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq pelo auxílio financeiro para a realização desta pesquisa.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

GLOSSÁRIO E ABREVIATURAS

API - *Application Programming Interface* (Interface de Programação da Aplicação)
ARPA - *Advanced Research Projects Agency*. Agência norte-americana que fomenta pesquisas em áreas avançadas.
ASR - *automatic-speech recognition* (reconhecedor automático de fala)
BBS - *Bulletin Board System* (Sistema de Quadro-mural)
DETER - Departamento de Transportes e Terminais
DNER - Departamento Nacional de Estradas de Rodagem
DOS - *Disk Operating System*. Sistema operacional para microcomputadores padrão IBM-PC, desenvolvido pela Microsoft Co.
FFT - *Fast Fourier Transform* (Transformada Rápida de Fourier). Processo que transforma uma função no domínio do tempo em uma função no domínio da freqüência.
FURB - Universidade Regional de Blumenau
GETRAN - Grupo de Estudos em Transportes
GUI - *Graphical User Interface* (Interface gráfica para usuário)
LVQ - *Learning Vector Quantization* (Quantização do vetor de aprendizado)
MCI - *Media Control Interface* (Interface de controle de mídia)
Quantização - pode ser entendido como o processo de dividir algo em partes menores.
SGCIF - Sistema Gerenciador de Central de Informação de Fretes
UFSC - Universidade Federal de Santa Catarina

RESUMO

A necessidade da evolução das interfaces homem-máquina gerou uma série de pesquisas na área de reconhecimento de padrões, procurando tornar naturais estas interfaces. A tarefa de reconhecimento de voz, realizada por computadores digitais, vem sendo pesquisada e desenvolvida nas últimas décadas, buscando alcançar o modo mais natural de comunicação humana - a fala. Este trabalho procura demonstrar a viabilidade e potencialidade dos sistemas comandados por voz, construindo um protótipo de software capaz de responder em tempo real a um comando falado por um usuário. Ele se utiliza das técnicas de redes neurais artificiais para realizar o reconhecimento das palavras faladas. A interface de voz construída para reconhecer as palavras, que operam um protótipo do Sistema Gerenciador de Central de Informação de Fretes (SGCIF) em Windows, aplica o modelo de rede neural Kohonen, alcançando uma taxa média de acerto de 84,84% no reconhecimento.

ABSTRACT

The need of the evolution of human-machine interfaces originated many researches at the field of pattern recognition, aiming to become natural these interfaces. The task of voice recognition, realized by digital computers, has been researched and developed at the last decades, aiming to reach the most natural way of human communication - the speech. This work demonstrates the executability and potential of voice commanded systems, developing a software prototype capable to respond to a command spoken by a user in real time. The prototype applies neural networks techniques to perform the recognition of spoken words. The voice interface developed to recognize words, which operate a Windows prototype of *Sistema Gerenciador de Central de Informação de Fretes* (SGCIF), applies the Kohonen neural network model, reaching a accuracy of 84,84%.

1. INTRODUÇÃO

1.1 Introdução do trabalho

Softwares comandados por voz são uma aspiração cada vez mais próxima do mercado mundial. Embora já existam alguns disponíveis, eles ainda têm um custo muito elevado e necessitam de aperfeiçoamentos. Entretanto, com o desenvolvimento de várias técnicas, dentre elas as Redes Neuronais, e com o aumento da capacidade de processamento dos computadores, brevemente, esta aspiração fará parte de interfaces amplamente difundidas e utilizadas, como ocorre hoje com as interfaces gráficas para o usuário (*graphical user interface: GUI*) - *Windows* ou *Presentation Manager* do OS/2.

Vários pesquisadores estão voltados à implementação de interfaces em linguagem natural. Com tais interfaces, o usuário não mais terá que se adaptar aos comandos do computador, caberá ao computador entender a linguagem comum do usuário. A linguagem poderá ser transmitida ao computador basicamente de duas formas: a escrita, utilizando-se de um teclado ou algum dispositivo de reconhecimento de caracteres; e a falada, suportada por dispositivos de reconhecimento de voz.

Este tipo de interface vislumbra a possibilidade de realmente ocorrer um diálogo falado entre o homem e a máquina, através de módulos que reconheçam a voz do usuário como entrada para a máquina e de módulos que sintetizem as informações de saída em sons, semelhantes aos da voz humana. Entretanto, as primeiras aplicações de larga escala, que surgem comercialmente, limitam-se ao reconhecimento de alguns comandos vocais.

Atualmente, grandes esforços na área de reconhecimento de voz são feitos com base em uma das técnicas emergentes da área de Inteligência Artificial: as Redes Neurais. Estas "redes" baseiam-se no funcionamento do próprio cérebro humano. Assim, possuem elementos de processamento que são modelos simplificados dos neurônios biológicos, e que simulam o funcionamento paralelo destes, em nosso cérebro.

As redes neurais[1] são utilizadas para realizar tarefas nas quais a performance humana é superior (melhor e mais rápida) a dos computadores programados com técnicas tradicionais. Exemplos de tarefas deste tipo são as de reconhecimento de padrões, como reconhecimento de imagens, caracteres escritos e voz.

1.2 Origem do trabalho

Esta pesquisa surgiu como Trabalho de Conclusão de Curso em Ciências da Computação - Bacharelado, na Universidade Regional de Blumenau - FURB, onde foi construído um protótipo para reconhecimento de duas palavras faladas.

O protótipo utilizava-se de um dispositivo digitalizador de sons (CI-500) ligado à saída paralela de um microcomputador padrão IBM-PC. Quando as palavras "Esquerda" ou "Direita" eram faladas ao microfone, o protótipo procurava fazer o reconhecimento através de uma rede neural Perceptron, que ao identificar o comando, movimentava para a posição correspondente um carro desenhado na tela do microcomputador.

Os resultados alcançados foram considerados bons, pois com uma rede neural de reduzidos recursos conseguiu-se alcançar uma taxa de acerto de 90%, quando utilizada pelo mesmo locutor que efetuou o treinamento da rede [HUG92].

1.3 Importância da pesquisa

"É abril de 1993 e você acaba de comprar um computador pessoal de última geração da Big Apple. Você abre a caixa e começa a montá-lo, mas para sua surpresa, você não vê um teclado, somente um pad especial e uma caneta que parecem um portfólio executivo. A caixa também contém roupas: uma blusa, um par de luvas e uma testeira.

---

[1] O termo neural também é encontrado como neuronal na literatura em português.

Você nota que não há monitor, somente um capacete e óculos. O que está acontecendo ?

"O que você está experimentando é o seu primeiro contato com a computação natural: fazer com que os computadores interajam com os usuários de uma maneira humana" [CAU92].

A totalidade deste "sonho" certamente está mais distante que o passado abril de 1993. Porém em todas as áreas citadas, cientistas e pesquisadores vêm obtendo sucessos na tentativa de aprimorar as Entradas/Saídas naturais (*natural I/O*).

Nos primeiros sistemas operacionais para computadores, as interfaces eram muito pouco amigáveis, uma vez que o operador deveria conhecer (memorizar) todos os comandos desejados, bem como sua sintaxe. A evolução destas interfaces chegou ao ponto atual, onde vários sistemas operacionais (Windows, OS/2, etc) são manuseados através de janelas gráficas, que, por intermédio dos ícones, mostram aos usuários as opções existentes, fazendo-os navegar espontaneamente por entre as possibilidades do sistema. O futuro certamente reserva uma maior evolução, vislumbrada hoje através das interfaces naturais.

Obviamente, o termo Entradas/Saídas naturais deve suscitar logo a idéia de uma interface de fala - o modo mais natural de comunicação - pois a maior parte da comunicação lingüística humana ocorre como fala [RIC88]. Esta interface natural de fala é, pois, algo que compreenda o que o usuário diz e que também lhe responda.

Nesta área, várias são as facilidades já disponíveis, não apenas em laboratórios, mas também no mercado. No início de 1992, a empresa francesa DECICOM lançou no mercado o *Vox'Scrib Base*, uma máquina portátil capaz de escutar um interlocutor humano, de lhe fazer perguntas e de responder com sua voz sintética. Ela pode, após um aprendizado de vinte minutos cada, reconhecer a voz de várias pessoas. Esta interface está disponível para MS-DOS a um preço de 45.000 F (em torno de US$ 9700) [BEL92].

Em Outubro de 92, a Microsoft lançou a *Windows Sound System*, uma placa de som para o ambiente Windows 3.1, que custava US$289. Os mais fortes atrativos da *Windows Sound System* são os recursos de voz, entre eles a possibilidade de o usuário acionar as opções dos menus através de comandos vocalizados, dispensando o uso do mouse e do teclado. Atualmente este pacote está na versão 2.0, porém

segundo [QUA94] a navegação dirigida por voz através do Windows continua muito primitiva.

A *Digital Equipment Corp.* gastou em torno de US$500.000 para equipar os barcos da Fundação América3 para participarem das regatas da *America's Cup* de 1992. O hardware e software envolvidos apresentavam reconhecedores de fala contínua como interface para computadores a bordo, executando programas de gerenciamento de regatas. Utilizando o *Verbex Voice Systems* de Edison, Nova Jersey, os engenheiros da Digital implementaram uma interface mais "marinheiramente" amigável. Todos esses avanços resultaram em um sistema mais rápido, fácil e seguro, facilitando as funções do navegador [INS92].

Para ilustrar, ainda poderiam ser citados alguns sistemas de resposta interativa por voz - sistemas que provêem informações de uma base de dados através de fala sintetizada e reconhecimento de voz das perguntas de um usuário, normalmente por telefone - alguns simples e baratos como o *Speech Master* da Speech Soft Inc (US$895) ou o *Computerfone III* da Suncoast Systems Inc (US$695-995), ou mais sofisticados e caros, como o *Voice Information Processing Server* da Octel Communications Corp. (US$13.500 - 588.000) ou ainda o *Audio Info Engine* (AIE) da Gralin Associates Inc (US$500.000 - US$1 milhão) [JEN91].

Estas interfaces têm seu poder de reconhecimento e compreensão de voz um tanto limitado e, contrariamente, um custo muito elevado. Outro fator importante a destacar é que uma interface construída para reconhecer os sons (fonemas) e a estrutura de uma determinada língua, não irá reconhecer os de outra, e este aspecto é bem mais complexo que uma simples troca de dicionários.

A síntese da fala, utilizada como meio de saída para as informações, já vem sendo utilizada há algum tempo, como por exemplo para enviar uma correspondência através do telefone (como frases faladas). O reconhecimento da fala, porém, teve menos utilização até então porque ainda existem pesquisas em busca de métodos apropriados, que possam prover um grau de funcionalidade aceitável a algoritmos de reconhecimento da fala [ALL90].

Assim, se interfaces de voz vêm sendo desenvolvidas e aprimoradas como um novo e fácil meio de Entrada e Saída de informações em um computador, é de extrema importância o domínio da técnica de reconhecimento de voz, pois logo ela se tornará um fator

essencial para a competitividade de qualquer software no mercado mundial.

1.4 Objetivos da pesquisa

O trabalho demonstra a viabilidade e potencialidade dos sistemas comandados por voz, construindo um protótipo de software capaz de responder em tempo real a um comando falado por um usuário.

O protótipo utiliza como aplicação demonstrativa o Sistema Gerenciador de Central de Informação de Fretes (SGCIF), na sua versão 2.0, convertido para o ambiente *Microsoft Windows 3.1*. Um usuário poderá, portanto, operar os menus do SGCIF através da voz.

O reconhecimento de voz será efetuado com base em técnicas de Inteligência Artificial, particularmente, em Redes Neurais Artificiais.

Como objetivos específicos deste trabalho tem-se:

1. aplicar o modelo de rede neural Kohonen para a tarefa de reconhecimento de padrões de fala;
2. testar a viabilidade da utilização de diferentes formas de pré-processamento dos sinais de fala digitalizados;
3. construir a interface de reconhecimento de voz no ambiente *Microsoft Windows 3.1*.

1.5 Organização da dissertação

Este trabalho é composto de cinco capítulos que apresentam o tema de reconhecimento de voz, suas implicações e soluções.

O Capítulo 2 - Interfaces Naturais - descreve o que se espera do futuro das interfaces, relatando os vários modos de entrada e saída esperados para os próximos anos e como irão interagir na estação de trabalho do futuro. Situa também o uso de interfaces de voz, relatando o histórico de pesquisas nesta área.

No Capítulo 3 - Redes Neurais e Reconhecimento de Palavras Faladas - são comentadas as vantagens da utilização de redes neurais nos problemas de reconhecimento de padrões, onde se destaca o reconhecimento de fala.

O Capítulo 4 - Aplicação - destaca os esforços realizados para se alcançar os objetivos propostos, explicando a aplicação demonstrativa (SGCIF) e relatando os estágios desenvolvidos na construção do protótipo. Informações a respeito de detalhes de construção e os resultados obtidos são apresentados neste capítulo.

O Capítulo 5 - Conclusão - encerra o trabalho relatando as conclusões e sugestões para futuros trabalhos.

2. INTERFACES NATURAIS

2.1 Introdução

Para que se possa compreender melhor como se processa a interação entre o homem e o computador, deve-se procurar entender o processo da comunicação humana.

2.1.1 Comunicação e Linguagem

As pessoas se comunicam para comandar, interrogar, responder, prometer, convencer as outras pessoas. Todas as pessoas quando têm uma idéia e pretendem compartilhá-la com outras - não importando seu objetivo - utilizam-se da comunicação. Comunicar é compartilhar um modelo, torná-lo comum [FIS87].

Para que haja a comunicação é necessário um veículo, uma forma de as duas entidades comunicantes compartilharem o mesmo modelo. "Uma linguagem é um conjunto de signos e símbolos que permitem um grupo social de se comunicar e facilita o pensamento e as ações dos indivíduos" [FIS87]. Thro [THR91] define linguagem como sendo "um modo de comunicação estruturado e inteligente que é falado, escrito ou passado por sinais."

Os seres humanos utilizam-se principalmente da linguagem natural para se comunicarem. Para Savadovsky [SAV88], "a linguagem natural é uma das formas mais humanas de manifestação externa da atividade mental ... Em particular, a linguagem é um meio muito rico para comunicação entre as pessoas." Linguagem natural é a comunicação estruturada e inteligente entre pessoas. Ela consiste de sons organizados; vocabulário; estruturas, tais como alfabetos ou outras representações simbólicas; gramática ou sintaxe; significado dependente da estrutura ou semântica; e métodos de interpretação daquilo que é ouvido ou lido [THR91].

2.1.2 Interface homem-computador

Além de um veículo para a comunicação, as duas entidades devem possuir meios de se comunicar. Em um sentido amplo, uma interface é um dispositivo que serve de limite comum a várias entidades comunicantes, as quais se exprimem em uma linguagem específica a cada uma. Para que a comunicação seja possível, o dispositivo deve assegurar a conexão física entre as entidades e efetuar as operações de tradução entre os formalismos existentes em cada linguagem. Uma vez que a comunicação esteja estabelecida, a interação (ação recíproca) pode ocorrer entre as entidades [COU90]. Para Thro [THR91], uma interface é um local para encontro ou interação.

No caso da interface homem-computador, a conexão entre estas duas entidades se realiza entre a imagem do sistema (ou seja, sua manifestação externa) e os órgãos sensoriais-motores do usuário; a tradução se efetua entre os formalismos do sistema e os do usuário [COU90].

Barthet [BAR88] afirma que há uma linguagem de interação permitindo ao usuário, por meio de um vocabulário e de uma sintaxe, expressar as operações que ele deseja efetuar à máquina. Estas operações manipulam comandos ou dados.

O vocabulário desta linguagem de interação pode ser formado de diferentes modos:

- códigos numéricos;
- códigos mnemônicos;
- palavras da língua portuguesa;
- palavras desconhecidas;
- pictogramas.

Porém, segundo Scapin [apud BAR88], a utilização de palavras da língua natural do usuário (no caso, a língua portuguesa) facilita a interação, pois ele utilizará uma linguagem mais intuitiva.

2.2 Interface em Linguagem Natural

2.2.1 Diálogo homem-máquina

A língua, seja escrita ou falada, é o meio de comunicação entre o homem e a máquina [PIE87]. Até os anos 70, tinha-se um "mosteiro" de especialistas em informática para manipular o computador; homens que dominavam a utilização de linguagens especializadas/especificas no uso do computador. Na década de 80, com o advento e expansão de microcomputadores e suas redes, o usuário final foi colocado frente-a-frente com a máquina e seus dados. Desta forma, linguagens mais naturais, bem como meios de acesso mais naturais eram prementes.

Em adição ao exposto, Feingenbaum coloca que "a riqueza das nações será amanhã essencialmente a informação e o conhecimento" [apud PIE87]. Esta informação e conhecimento tem como suporte (veiculo) básico nada mais que linguagem natural, visto que grande parte do conhecimento da humanidade encontra-se armazenado (geralmente escrito) nesta forma. Daí advém a necessidade de compreender esta linguagem, seja de forma escrita ou oral.

Pierrel [PIE87] e Anick [ANI93] entendem que com o crescimento de tamanho e importância das bases de dados, os computadores necessitam ter meios de interpretar a linguagem natural. Desta forma um usuário pode mais facilmente encontrar algum argumento de pesquisa na base de dados, sem se preocupar com sua especificação exata - seja em termos de comandos de busca, seja em palavras a pesquisar - permitindo-lhe realizar suas pesquisas através do vocabulário que lhe é conhecido, sendo o computador responsável por oferecer-lhe sinônimos e ajuda direcionada ao argumento.

Scapin [apud BAR88] afirma que vários profissionais de informática estimam que a utilização da linguagem natural é o que melhor pode se oferecer ao usuário em termos de interface. Contudo para haver comunicação é necessário haver diálogo, senão se reduz a uma simples transmissão de informações que não satisfaz ao ser humano [PIE87]. Um exemplo desta situação é o comportamento de um correspondente ao telefone face uma secretária eletrônica. Com exceção das pessoas acostumadas a este tipo de aparelho, o correspondente se encontra despreparado, ao ponto de esquecer às vezes a mensagem que gostaria de transmitir ou de tentar provocar um diálogo para facilitar sua comunicação. Porém o diálogo ocorre

normalmente por comunicação em multicanais [PIE87], ou seja, utilização da fala, de gestos, de olhares, etc. Fischler e Firschein [FIS87] denominam de linguagem do corpo (*body language*) todos os componentes não-verbais que compõem um diálogo, mas que podem afetar o sentido da comunicação pelo canal verbal.

2.2.2 Entradas/Saídas Naturais

A utilização de linguagem natural não garante que a interface seja natural. Isto é, fazer com que o usuário possa digitar seus comandos de acordo com seu vocabulário coloquial facilita seu acesso ao computador, porém oferecer-lhe uma interface através da qual ele consiga dar entrada a esta mesma linguagem por voz ou escrita manual (ou ambos) seria mais próximo ao modo comum dele comunicar-se.

Crane e Rtischev [CRA93] apontam que a utilização de reconhecimento de voz e de escrita manual derrubarão as barreiras que teclados, mouses e GUIs impõem à comunicação natural com o computador. Vários fabricantes de sistemas dizem que pode se esperar para ver computadores combinando alguma forma de entrada por voz e caneta dentro de poucos anos. Raj Reddy, diretor da *School of Computer Science* da Universidade de Carnegie Mellon e pesquisador há mais de 30 anos na área de fala, prediz que computadores pessoais utilizarão entradas por voz e caneta dentro de cinco anos [CRA93]. E com a miniaturização cada vez mais acentuada dos computadores, será possível pensar em computadores que podem ser usados como roupas ou chapéus, ou seja, tão pequenos e leves que são facilmente transportados e com entrada e saída por voz, através de um microfone direcional e de pequenos fones de ouvido, por exemplo.

Cada método de comunicação com um computador tem suas vantagens e desvantagens [TEB95]. Um mouse ou joystick não pode ser facilmente usado para digitação e um teclado ou a fala não são ideais para manobrar um cursor na tela. Muitos dispositivos podem ser associados a controle além de suas funções principais de posicionamento ou entrada de dados. Rudnicky [RUD93] apresenta o seguinte quadro (Quadro 1) que aloca tarefas a modos de entrada. Alguns modos de entrada atendem atividades particulares melhor que outros, conforme as indicações: B representa que o modo de entrada é uma boa escolha para a atividade; A representa que o modo de entrada é meramente adequado; e I representa que o modo de entrada é inadequado.

Atividade	Fala	Caneta	Teclado	Apontador/Mouse
Assinalar	I	B	A	A
Ditar, anotar	B	A	A	I
Verificar o usuário	B	B	A	I
Criar gráficos	I	B	A	B
Preencher formas/figuras	B	B	B	B
Conferir listas	A	B	A	B
Comando e controle, comunicações e redes	B	A	A	A
Computação por planilhas e financeira	A	B	B	A
Agendar, planejar e organizar	B	B	A	A

Quadro 1: Atividades versus modos de entrada [RUD93]

"Caneta, mouse, teclado e voz coexistirão no *desktop* do futuro" prediz Lempesis [MEZ93], afirmando que nenhuma opção de entrada dominará. O usuário selecionará seu dispositivo de entrada baseado na aplicação e nas suas preferências pessoais. Tebbutt [TEB95] vislumbra a utilização de várias opções de entrada, como reconhecimento de voz, canetas, telas sensíveis ao toque e outros estranhos aparelhos de realidade virtual, porém ele crê que nenhum se sobressairá como único, sendo a combinação de vários deles a forma de entrada em um futuro próximo.

Meisel [MEI93] afirma que um microfone conectado a um computador pessoal pode possuir várias finalidades, entre elas a de armazenar recados, a de verificar a identidade do locutor para fins de segurança e a de reconhecer palavras na fala. Destas funções, o reconhecimento de voz é a que tem maior potencial para fundamentalmente mudar o modo de interação com o computador: "A tecnologia está produzindo uma nova interface homem-computador".

"O objetivo da interação homem-computador é, ou deveria ser, prover uma interface tão natural quanto possível. De fato, a solução perfeita seria aquela na qual o usuário nem percebesse a utilização de uma 'interface'... Então, talvez a falta de uma interface seja o Nirvana dos usuários de computador. ", expõe Tebbutt [TEB95].

Outros modos de interação homem-máquina podem ainda ser listados, como telas sensíveis ao toque, luvas (*datagloves*), sistemas de câmeras que captam gestos, sistemas que controlam o

movimento dos olhos, etc. Todos estes podem ser formas de entrada/saída que proveriam os multicanais citados por Pierrel e a linguagem do corpo citada por Fischler e Firschein, no tópico 2.2.1.

2.3 Interfaces de voz

2.3.1 Introdução

As línguas naturais são principalmente faladas [MAI85]. Uma das razões históricas para a predominância da fala é que, visto ser o homem um animal que trabalha, é vantajoso utilizar a audição para a comunicação, deixando os demais sentidos livres para exercerem outras atividades, que podem ou não ter funções comunicativas [MAI85]. Rich [RIC88] ainda destaca que a linguagem escrita é uma invenção recente e ainda desempenha um papel menos crítico que a fala, na maioria das atividades.

Focalizando, então, a linguagem falada, destacam-se duas vantagens na sua utilização em um sistema de comunicação homem/máquina [PIE87]:

- a fala é indispensável , em algumas situações, para substituir outros canais de comunicação, como a relação piloto-avião ou para deficientes físicos;
- a fala pode ser mais eficaz que outros modos de comunicação. Em comparação à comunicação escrita, a fala permite melhores performances em questões de tempo de resolução de problemas e emissão de mensagens.

Allen [ALL90] afirma que a utilização da fala humana como meio de entrada/saída expandiria enormemente o uso de sistemas em computador. Ele explica que os sistemas têm a tendência de se tornarem cada vez mais complexos, concluindo que, facilitando a interação homem-máquina, haveria uma maior utilização destes sistemas. "Reconhecimento de fala é um dos pontos-chave do cliente de negócios" diz Bob McBreen, gerente de produto para a Microsoft *Windows Sound System*. Ele acredita que o reconhecimento de fala será parte integrante da computação no futuro [MEI93].

2.3.2 Abordagens de reconhecimento de fala

Pierrel [PIE87] apresenta duas grandes abordagens para o reconhecimento. Estas abordagens não representam uma classificação rígida, podendo até ser complementares entre si a fim de solucionar o problema de reconhecimento.

- Método global e método analítico

O método global, também conhecido por reconhecimento de palavras, utiliza basicamente as técnicas de reconhecimento de formas para comparar, globalmente, a palavra a reconhecer das diversas formas de referência armazenadas.

O tratamento acústico preliminar é bem simples: a mensagem a identificar é considerada como uma forma atômica, sem problemas de segmentação. O ponto delicado é a melhor forma de representação da palavra.

Esta abordagem global se revela insuficiente se pretendesse tratar com grandes vocabulários ou fala contínua. Se faz, então, necessário adotar a abordagem analítica, que consiste em segmentar a mensagem em constituintes elementares (fonemas, meia-sílabas, sílabas, etc). Após identificados estes últimos, reconstituir enfim a frase pronunciada por etapas sucessivas: léxica, sintática, etc.

No quadro 2 são observadas aplicações dos dois métodos em relação à entrada da fala.

	Palavras isoladas	Fala contínua
Global	Reconhecimento de palavras (pequeno vocabulário) Sistemas comercializados	Localização de palavras dentro de frases
Analítica	Reconhecimento de palavras (grandes vocabulários)	Localização de palavras Reconhecimento e compreensão de frases

Quadro 2: Diferentes tipos de reconhecimento [PIE87]

- Reconhecimento x Compreensão

Uma segunda classificação foi proposta pelos pesquisadores do projeto ARPA, que introduziram o termo Compreensão da fala (*speech*

understanding) em oposição ao termo Reconhecimento de fala (*speech recognition*).

Reconhecimento de fala consiste no reconhecimento de fonemas, sílabas, palavras para formar a mensagem original, como foi pronunciada. Como exemplo deste tipo tem-se as máquinas de ditar e editores de texto por fala.

Compreensão de fala baseia-se no entendimento do senso, do significado da mensagem, visando fazer com que o sistema execute algo. Para tal, são aceitos eventuais erros. Utilizando o exemplo de Verhaeghe [VER92]: em uma aplicação Windows, poderia ser falado "*Please start ...euh... spreadsheet*". O comando é reconhecido como "*start application spreadsheet*" ignorando algumas palavras (*please*) e subentendendo outras (*application*), pois o entendimento se dá por certas palavras-chave (*start spreadsheet*).

Esta distinção foi feita pois notou-se que a habilidade de um sistema em responder inteligentemente à fala era um critério mais significativo para a avaliação de sistemas de fala. Além disso, acreditou-se que o sinal de fala era uma pobre fonte de informação e que o conhecimento do contexto de uma pronúncia era essencial para seu reconhecimento e interpretação com sucesso [ALL90].

2.3.3 Reconhecedor automático de fala

Antes de se conhecer um sistema computadorizado de reconhecimento de fala, far-se-á uma verificação superficial de como isto ocorre naturalmente, ou seja, em seres humanos.

Para que haja comunicação é necessária a existência de dois personagens: o locutor e seu interlocutor, ou ainda, o emissor e o receptor da mensagem.

O emissor produzirá uma mensagem fazendo com que determinada idéia que possua seja transformada em sons, através do comando de nervos motores do aparelho fonador, ou seja, o emissor tem uma idéia e codifica-a em símbolos que são transmitidos ao receptor [THO81].

O receptor decodifica estes símbolos em um *código interno* (idéia), ou seja, o receptor perceberá a mensagem, através de nervos sensoriais do seu aparelho auditivo, procurando transformar os sons recebidos na idéia original. A comunicação pode ser

considerada boa caso haja um isomorfismo entre os estados internos de idéia tanto do emissor quanto do receptor [THO81].

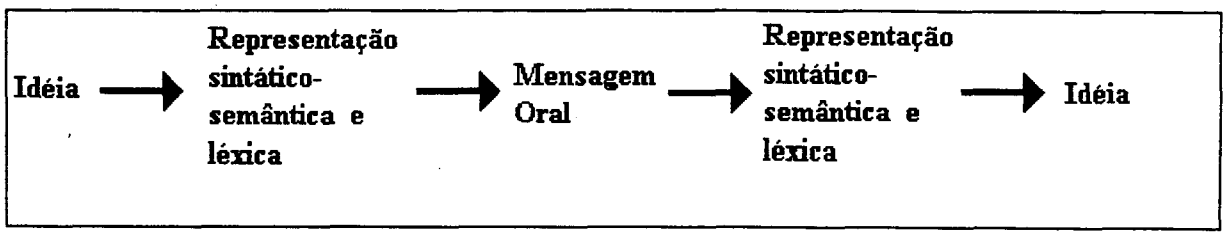

Figura 1: Codificação e decodificação sucessivas na comunicação oral [PIE87].

Um reconhecedor automático de fala (*automatic-speech-recognition: ASR*) será sempre o receptor da mensagem. Ele fará a percepção das ondas sonoras da mensagem e executará algum processamento procurando "captar a idéia" do emissor.

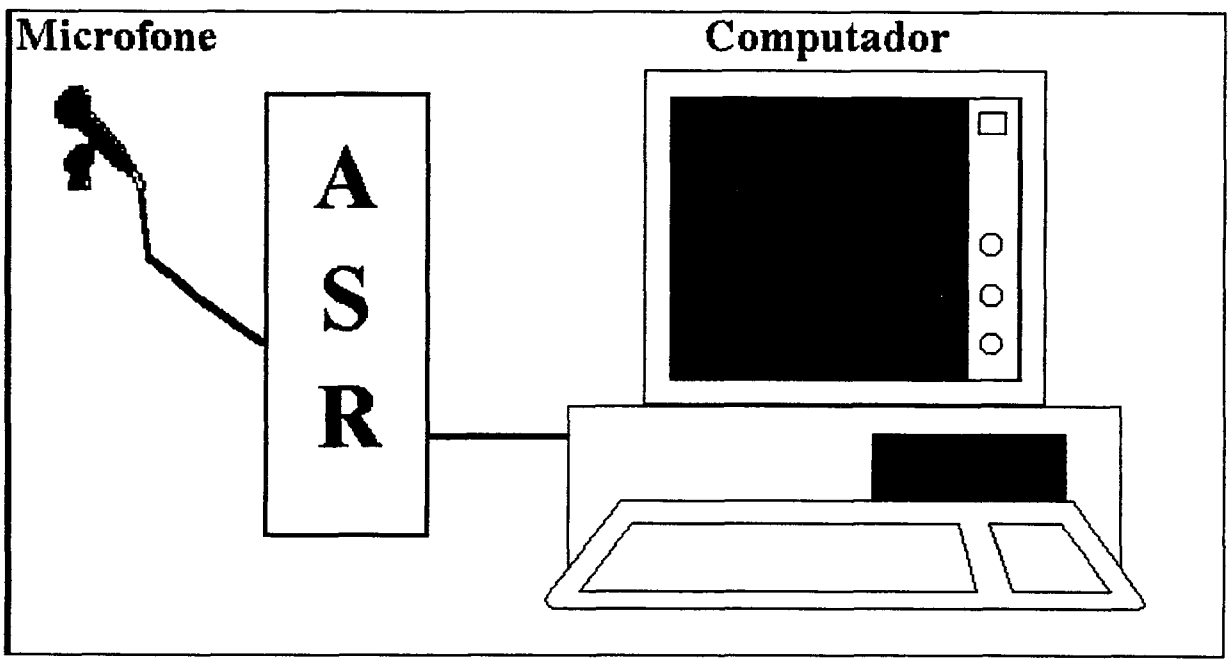

Figura 2: Configuração típica de um ASR [DAS92].

Uma configuração típica de um ASR pode ser vista na figura 2, onde há um microfone (fazendo as funções dos nervos sensoriais do aparelho auditivo) conectado a um sistema ou dispositivo (hardware+software) ligado ao computador.

Este dispositivo consiste basicamente em:

- hardware capaz de transformar as ondas sonoras captadas (sinal analógico) em números (sinal digital) para algum processamento pelo software;

• software que capte a massa de dados numéricos enviada pelo hardware, e reconheça alguma palavra ou execute algum comando.

No desenvolvimento e aprimoramento deste software é que se concentra o esforço de milhares de pesquisadores em todo o mundo. Em linhas gerais, pode se dizer que cabe a este software [ALL90] [BRI90]:

• análise do sinal e extração de parâmetros: a quantidade de bits por segundo gira na faixa de 50.000 nos sistemas com boa qualidade de recepção. Esta é uma massa muito grande de dados para ser tratada. Logo, devem ser aplicados meios de redução ou extração desta informação, sem perder as características do sinal representativo da mensagem. Para tal, são utilizadas várias técnicas, entre elas: transformada discreta de Fourier, banco de filtros, densidade de passagem por zero, etc;

• determinação do ponto-final da fala: determinar quando os dados recebidos não significam mais a fala propriamente dita, mas apenas ruído do ambiente, ou mesmo o silêncio;

• normalização em freqüência e tempo: a necessidade de normalização em freqüência origina-se do fato de que cada usuário-locutor possui um timbre e entonação diferentes, fazendo com que as freqüências para um mesmo fonema sejam diferentes. A normalização em tempo advém das diferentes velocidades com que pode ser dita uma palavra;

• reconhecimento: após terem sido efetuados os passos anteriores (na ordem apresentada ou não), algum modo de identificação deve ser aplicado. Várias técnicas são utilizadas, como: comparações com dicionários, regras de produção, programação dinâmica, modelo escondido de Markov (*hidden Markov model : HMM*), e mais recentemente, redes neuronais.

2.3.4 Aplicações do reconhecimento de fala

A imaginação é o limite para definir aplicações para este novo tipo de interface. Contudo, há que ser lembrado as atuais restrições existentes para estas aplicações. A primeira e mais importante delas é a capacidade: de memória e de processamento. As técnicas atualmente utilizadas não podem evoluir muito mais em razão das restrições de capacidade. Isso acarreta em soluções hoje existentes que possuem contexto bem definido, ou seja, área de aplicação bem definida com vocabulário específico deste domínio, tendo como conseqüência uma pequena quantidade de palavras, para que possam ser obtidos resultados em tempo real.

Pierrel [PIE87] coloca que é vantajoso restringir o universo do diálogo, pois desta forma não será necessário tanto conhecimento a respeito de sintaxe, semântica, prosódia, etc. diminuindo a possibilidade de ambigüidades e simplificando o processo de coleta de informações.

Fischler e Firschein [FIS87] lembram que normalmente o vocabulário utilizado pelas pessoas é bem menor do que se pode imaginar. Os quadros 3 e 4 demonstram isto. Eles são para o inglês, porém são semelhantes para português, francês ou qualquer outra lingua natural. O quadro 3 mostra que o vocabulário normal de um adulto consiste de cerca de 10% das palavras de um dicionário resumido.

Fonte	Número de palavras
Criança	3.600
14 anos de idade	9.000
Adulto	12.000-14.000
Dicionário resumido	150.000
Divina Comédia de Dante	5.900
Poemas de Homer	9.000
Trabalhos de Shakespeare	15.000-25.000

Quadro 3: Tamanho do vocabulário empregado por várias fontes. Kendratov [apud FIS87].

No quadro 4 pode-se observar que com um vocabulário de 3000 palavras, consegue-se reconhecer 90% das palavras em um página de texto comum. Com um vocabulário de 1000 palavras consegue-se o mesmo índice de reconhecimento para palavras faladas.

Vocabulário da linguagem falada	Vocabulário da linguagem escrita	Probabilidade de ocorrência de palavras na fala ou texto
750	--	75,0%
--	1000	80,5%
--	2000	86,0%
1000	3000	90,0%
2000	5000	93,5%

Quadro 4: Freqüência do uso de palavras na linguagem escrita e oral. Kendratov [apud FIS87].

Os sistemas atualmente existentes trabalham com uma quantidade que varia de algumas palavras, passando por centenas, chegando até 50.000 nos sistemas mais sofisticados [MEI93].

O Tangora, da IBM, é um sistema tipo máquina de escrever comandada por voz, o qual trabalha com as 20.000 palavras mais comuns em ambiente de escritório. Contudo possui uma outra restrição: o ditado deve ocorrer de forma mais pausada. Esta tecnologia é conhecida como entrada de fala isolada ou discreta, em contraposição à entrada de fala contínua. Porém Tebbutt [TEB95] lembra que o teclado padrão QWERTY foi projetado para que as pessoas não conseguissem datilografar muito rapidamente, o que travaria o sistema mecânico das máquinas de escrever da época. Desta forma, a pequena pausa, porém não-natural, realizada entre as palavras durante o ditado para um sistema de reconhecimento de fala pode ser encarada como normal, visto que a demanda da tecnologia tem precedentes sobre os desejos do usuário.

Mesmo obedecendo às restrições comentadas, várias aplicações poderiam ser destacadas:

- em linhas de produção para descartar peças falhadas através de comandos vocálicos para movimentar a esteira de produção;
- na mesma idéia de movimentar esteira, manuseio de bagagens aéreas (de mão ou não);
- em situações mãos e olhos ocupados , como por exemplo, para o motorista discar um telefone em um carro em movimento, ou um piloto dentro da cabine de um avião;
- auxílio a deficientes físicos;
- em caixas bancários automáticos, para solicitar saldos, extratos, etc;
- operações por telefone, como consultas de saldos, de notas em colégio, etc
- preparo de relatórios médicos ou odontológicos durante ou após a consulta.

2.4 Histórico dos trabalhos em reconhecimento de fala

Pierrel [PIE87] relata que contrariamente à síntese de fala, cujos primeiros trabalhos datam do século XVII, foi preciso esperar

o meio deste século para aparecerem as primeiras experiências em reconhecimento.

Conforme Pierrel [PIE87], por volta de 1930, o americano R.J.Wensley construiu o TeleVox, primeiro autômato capaz de receber ordens por telefone e executar alguns movimentos correspondentes.

Os primeiros sistemas realmente de reconhecimento de voz aparecem somente em 1950: Daves, em 1952, apresenta um sistema inteiramente de cabos capaz de reconhecer os dez números pronunciados por um locutor. Sistema este aperfeiçoado em 1958 para uma versão que aceita diversos locutores; Olson e Belar, em 1956, propuseram um sistema ambiciosamente chamado de máquina de escrever fonética, capaz também de reconhecer uma dezena de palavras; Denes, em 1958, define um sistema em duas etapas, no qual a primeira etapa realiza um reconhecimento puramente acústico que, na segunda, será refinado pela utilização de conhecimentos lingüísticos.

Em 1960, a aparição dos métodos numéricos e a utilização do computador dão uma nova dimensão a estas pesquisas. Em 1966, sistemas em laboratório conseguem identificar corretamente de trinta a cinqüenta palavras ditas por diferentes pessoas. Estas experiências eram baseadas na comparação das formas das palavras. No início dos anos 70, a programação dinâmica otimiza este tipo de abordagem.

Em 1968 um importante passo é alcançado, Alter e Reddy verificam a utilidade das informações lingüísticas no reconhecimento da fala. Contudo, os primeiros trabalhos a utilizarem esta realização são os de Vicens, em 1969, e Tubach, em 1970. Estas pesquisas sobre sistemas de reconhecimento de palavras isoladas continuam a progredir em variados eixos.

Paralelamente surgem as primeiras investigações no tratamento de fala contínua, que vão seguindo fracamente até 1973, quando é publicado o relatório Newell, fortemente influenciado pela Inteligência Artificial, para relançar estes estudos no quadro de um projeto americano financiado pela ARPA (*Advanced Research Projects Agency*) de 1971 a 1976. Os objetivos deste projeto definiam que o sistema deveria aceitar a fala contínua, aceitar um grande número de locutores cooperativos, trabalhar num ambiente calmo, com um bom microfone, compreender um vocabulário de mil palavras, utilizar uma sintaxe artificial no escopo de uma tarefa precisa, responder em tempo próximo ao real (*a few times real time*) em uma máquina de cem MIPS (milhões de instruções por segundo).

Este projeto faz surgir várias pesquisas em reconhecimento de fala contínua. Muitos sistemas foram propostos. Julgado positivo, o projeto ARPA serviu de catalisador de pesquisadores neste domínio [PIE87].

Sistemas como DRAGON (Baker em 1975), HEARSAY (Lesser, Fennel, Erman e Reddy em 1975) e HARPY (Lowerre em 1976) trabalhavam com um discurso contínuo de um único usuário e um vocabulário de até 1.000 palavras, obtendo taxas de acerto entre 84 e 97%. O TANGORA da IBM (citado no tópico 2.3.5) tem uma versão lançada na metade da década de 80, a qual sacrificava a fala contínua para atingir uma taxa de acerto de 97% para um vocabulário de 20.000 palavras. Um sistema criado nos Laboratórios Bell em 1987 para o reconhecimento de dígitos para números de telefone, também obteve 97% de precisão. O SPHINX (Lee e Hon em 1988) foi o primeiro sistema a alcançar alta precisão (96%) em fala contínua, independente do locutor e em tempo real com vocabulário de 1.000 palavras [RIC93].

Não apenas nos Estados Unidos estas pesquisas vêm sendo realizadas. Ao longo da década de 80, várias são as empresas que se lançaram em busca do reconhecimento de fala, bem como muitos centros de pesquisa. Pode-se citar, entre outras [PIE87]:

- EUA - IBM e SRI (*Stanford Research Institute*) ;
- Japão - NEC (*Nippon Electronic Company*), o projeto PIPS (*Pattern Information Processing Signal*), NTT (*Nippon Telegraph and Telephon Public Corporation*) e várias universidades;
- Europa - Alemanha, Itália, Inglaterra, França, etc.

No final da década de 80, Teuvo Kohonen da Universidade de Tecnologia de Helsinki desenvolveu uma máquina de escrever por voz utilizando a combinação de várias tecnologias disponíveis: processamento de sinal digital, sistemas baseado em regras e redes neuronais. A máquina obteve alguns bons resultados. Foi testada utilizando-se casos extremos de conversão fala-texto: vários locutores, fala contínua e grande vocabulário. Com um treinamento de apenas 100 palavras por locutor, a máquina chegou a uma taxa de 92 a 97% de acerto nas conversões, com *delay* de 1/4 de segundo de resposta. São bons resultados, mas não ainda suficientes para aplicações comerciais. Contudo, o casamento de várias tecnologias permitiu a visualização de um futuro promissor [CAU92].

Em 1994, vários artigos proclamam a chegada do reconhecimento de voz no mercado. Fritz [FRI94] escreve que o reconhecimento de fala finalmente ganha algum respeito, prevendo que até o ano de

1999 esta área deve movimentar uma cifra em torno de US$ 1 bilhão. E testes conduzidos pela Seybold Publications [SEY94] consideraram os sistemas *Personal Dictation System* da IBM para OS/2 e *DragonDictate* da Dragon Systems para MS-DOS como impressionantes, pois apesar dos problemas detectados quando da existência de sons semelhantes("*to many people*" e "*too many people*") a disponibilização dos resultados imediatamente permite ao usuário fazer qualquer correção necessária.

2.5 Considerações Finais

É cada vez maior a necessidade do homem interagir de forma mais natural com o computador, ou genericamente, com as máquinas. Modos de entrada/saída naturais vêm sendo vislumbrados para prover esta comunicação, que acarreta em várias facilidades e vantagens.

A fala é um dos meios de comunicação mais naturais entre os seres humanos. Assim, é de suma importância a aquisição e utilização de tecnologias para torná-la realidade nas interfaces de computador. Ao longo das últimas décadas vários trabalhos foram desenvolvidos, porém ainda não há a possibilidade de uma utilização plena de tal forma de interação.

O estudo apresentado neste capítulo demonstra que houve uma significativa evolução nos resultados de pesquisas desenvolvidas para alcançar estas tecnologias. O aumento da capacidade dos computadores - processamento e memória - contribuiu decisivamente para estes avanços.

Contudo, vale ressaltar que a comunicação humana não ocorre apenas através da fala e sim através de multicanais, formados pelo conjunto dos sentidos. Grandes avanços científicos e tecnológicos ainda estão reservados para estas áreas de pesquisa.

3. REDES NEURAIS E RECONHECIMENTO DE PALAVRAS FALADAS

3.1 Introdução

De acordo com Pierrel [PIE87], desde a Grécia Antiga até os dias atuais, toda a história da humanidade está ligada à retórica e eloqüência. Demóstenes, na Grécia, já dizia que "tudo dependia do povo e o povo dependia da fala". Hoje em dia, "os que têm a experiência da fala ocupam os postos-chave.", nota Bellenger [apud PIE87].

Grandes pesquisadores da psicologia, lingüística e psicolingüística, apesar de defenderem teorias e estudos às vezes um pouco divergentes, acreditam que o ser humano necessita de uma língua para pensar e raciocinar.

Piaget, que observou com paciência os fenômenos de aprendizado e de modelização nas crianças, acredita que há uma interação entre linguagem e processos de modelização, que são a base do que se costuma chamar de inteligência. Ele defende que as crianças possuem uma fala egocêntrica, contudo não via funções especiais nesta fala. Já Watson acreditava que tal fala, sob a pressão de não falar alto, era interiorizada para se tornar uma fala subvocal, equivalente ao pensamento. Vygotsky, que em suas pesquisas discordava dos dois autores anteriores, concluiu que: "Nossos resultados experimentais indicam que a função da fala egocêntrica é semelhante à da fala interior: ela não apenas acompanha a atividade da criança; serve de orientação mental, compreensão consciente; ajuda a vencer dificuldades; é fala para si mesma, ligada íntima e proveitosamente ao pensamento infantil. (...) No fim, ela se torna fala interior." [apud SLO80].

A fala não está limitada à emissão e recepção de sons, mas a toda uma estrutura sintática e semântica de uma língua na memória, ou seja, ao conhecimento de uma língua. Bellenger questiona: "O homem do fim do século XX não será capaz de pensar além de quando fala ?".

Deve-se lembrar também que o homem é um animal inteligente e o único que fala. Mas é inteligente porque fala? Não. Fala porque é inteligente. A fala é a exteriorização dos pensamentos.

Isto faz com que grande parte das pesquisas mundiais sobre compreensão e reconhecimento de fala estejam hoje computacionalmente baseadas na Inteligência Artificial. Se a fala é um processo cognitivo e inteligente do ser humano e a Inteligência Artificial visa extrair os conhecimentos necessários a por em prática as funções humanas, ou do cérebro humano, torna-se bastante natural que várias pesquisas de reconhecimento de fala se utilizem de uma das técnicas emergentes neste campo: as redes neurais.

3.2 Redes neurais aplicadas a reconhecimento de padrões

Notadamente redes neurais têm sido uma técnica utilizada nas tarefas de reconhecimento de padrões. Wang [WAN93] afirma que entre todos os domínios de utilização de modelos de redes neurais, o reconhecimento de padrões é o que possui maior potencial. Nos próximos tópicos, define-se a tarefa de reconhecimento de padrões e seu relacionamento com redes neurais artificiais.

3.2.1 Reconhecimento de padrões

Reconhecimento de padrões são tarefas que causam pouca dificuldade para os seres humanos, e até mesmo para os animais, contudo são um desafio confuso para a moderna tecnologia [WAN93].

O termo "reconhecimento de padrões" foi introduzido no início da década de 60 e originalmente significava a detecção de formas simples [KOH88]. Para Bezdek e Pal [BEZ92], há várias definições para o termo reconhecimento de padrões, porém aquela que mais o caracteriza é dada por Duda e Hart, em 1973, como sendo "campo interessado no reconhecimento por máquinas de regularidades significativas em ambientes ruidosos ou complexos", ou a procura por uma estrutura nos dados.

Há duas grandes motivações para estudos nesta área: a necessidade das pessoas em se comunicarem com máquinas computacionais através de linguagens naturais; e o interesse na

idéia de projetar e construir autômatos (máquinas inteligentes) que possam realizar certas tarefas com habilidades comparáveis à performance humana [BEZ92]. Estas tarefas envolvem percepção invariante em relação a equivalência de estímulos, posição, deslocamento, rotação, perspectiva, oclusão parcial, etc [IYE91].

Áreas de aplicação de reconhecimento de padrões incluem [BEZ92]:

- comunicação homem-máquina: reconhecimento automático de fala, reconhecimento da escrita, compreensão de fala, compreensão de imagens, processamento da linguagem natural;
- defesa: reconhecimento automático de alvos, orientação e controle;
- medicina: diagnose médica, análise de imagens, classificação de doenças;
- veículos: controladores de automóveis, aviões, trens, barcos;
- polícia e investigação: detecção criminal a partir da fala, escrita manual, impressões digitais, fotografias;
- estudo e estimativa de recursos naturais: agricultura, extrativismo, geologia, ambiente;
- indústria: CAD, CAM, teste e montagem de produtos, controle e inspeção de qualidade;
- sistemas domésticos: utensílios; e
- computadores: hardware e software difusos.

3.2.2 Técnicas para o reconhecimento de padrões

Algumas aplicações de reconhecimento podem requerer que seja encontrado uma ocorrência exata de um padrão, enquanto outras são satisfeitas por encontrar uma ocorrência aproximada. Esta crítica distinção separa os métodos basedos em abordagens simbólicas dos métodos basedos em universo de características e teoria de decisão estatística [FIS87]. Fu [FU83], Kanal e Dattatreya [KAN90] também denominam de abordagens sintáticas ou estruturais e de abordagens estatísticas ou numéricas.

Fu [FU82] é um dos autores que explora a idéia da representação **sintática** (ou estrutural) de um padrão para o problema de reconhecimento, utilizando gramáticas para esta tarefa. Ao definir uma gramática para descrever os padrões, o projetista estará definindo componentes primitivos (símbolos) e as regras de como estes componentes serão agrupados para formar os padrões desejados.

Outra corrente de reconhecimento de padrões é a do reconhecimento **numérico** de padrões que define que todo padrão pode ser representado por um vetor numérico, conhecido como vetor característico ou vetor do padrão. Este vetor é comparado a um vetor representativo de uma classe de padrões, também conhecido como vetor de similaridade ou de proximidade. O padrão será reconhecido ou classificado de acordo com a maior similaridade entre ele e o vetor representativo de uma classe. Muitas das técnicas aplicadas ao problema de reconhecimento através de estatística, lógica difusa e redes neurais estão nesta corrente [BEZ92].

As correntes do reconhecimento **contextual**, **conceitual** e **baseado em regras** são tentativas de adicionar contexto, conceito e conhecimento especializado em algum nível intermediário do processamento a fim de aproveitar a perícia humana nas tarefas de reconhecimento.

3.2.3 Aplicação de redes neurais

Técnicas e conceitos básicos a respeito de redes neurais artificiais têm sido amplamente cobertos pela literatura [LIP87] [KOH88]. Porém, cabe aqui destacar algumas vantagens do uso destas redes para que se possa melhor entender seu relacionamento e sua aplicação em processos de reconhecimento de padrões.

O reconhecimento de padrões é, por sua própria natureza, uma ciência não exata [BEZ92]. Enquanto alguns padrões podem ser identificados como bem estruturados ou estruturados adequadamente para serem definidos por uma gramática, como por exemplo a fabricação de um produto em uma linha de montagem, outros são de difícil modelagem ou difícil construção de uma gramática.

Fu [FU82] ao desenvolver a idéia da representação sintática para o problema de reconhecimento, descreve o processo de definição de uma linguagem para a descrição dos padrões, afirmando que não há soluções gerais e que a escolha é influenciada pela natureza dos dados disponíveis, da aplicação e da tecnologia. O projetista então vai realizar escolhas para a definição da gramática baseadas em suas experiência e vontade.

Esta natureza não estruturada do padrão a ser reconhecido torna o problema de reconhecimento difícil de ser tratado por paradigmas tradicionais de computação [IYE91]. Desta forma, o uso de redes neurais aparece como modo alternativo de resolução, pois segundo Iyengar e Kashyap [IYE91], ao invés de criar procedimentos lógicos, a construção destas redes envolve o entendimento informal do comportamento desejado para atender ao problema. Blum [BLU92] afirma que utilizando redes neurais há menor necessidade de se determinar, *a priori*, quais são os fatores determinantes sobre o modelo que se está desenvolvendo.

Bezdek e Pal [BEZ92] destacam quatro maiores vantagens do uso de redes neurais sobre muitas técnicas tradicionais de reconhecimento de padrões:

- adaptatividade: habilidade de se ajustar a novas informações;
- velocidade: via o paralelismo massivo;
- tolerância a falhas: capacidade de oferecer boas respostas mesmo com falta, confusão ou dados ruidosos;
- otimalidade: visto como taxa de erros em sistemas de classificação.

A capacidade de generalização da rede neural também se apresenta como vantagem no problema de reconhecimento de padrões. Tomando como exemplo o reconhecimento de escrita manual, pode-se imaginar a grande variedade de formas que uma mesma letra pode ser escrita, mesmo sendo por uma mesma pessoa. A inclinação, tamanho, pressão e traçado são algumas das variáveis que podem afetar o reconhecimento da escrita. A rede neural após aprender a distinguir alguns As de tamanho diferente de alguns Bs de tamanho diferente, será capaz de distinguir um A de qualquer tamanho de um B de qualquer tamanho. Desta forma, a capacidade de reconhecer padrões nunca antes vistos, porém semelhantes aos apresentados durante o treinamento, torna-se um diferencial perante muitas técnicas tradicionais, além de ajudar a superar ruídos indesejáveis nas entradas.

De modo geral, as redes neurais são um método de modelamento altamente recomendável para se lidar com sistemas abertos ou mais complexos, pouco entendidos e que não podem ser adequadamente descritos por um conjunto de regras ou equações.

3.3 Redes neurais aplicadas a reconhecimento de fala

3.3.1 Variáveis do problema de reconhecimento de fala

Na área de reconhecimento de padrões, o reconhecimento de fala sempre é destacado, figurando como um dos problemas que recebem grande atenção por parte da comunidade cientifica.

Ele envolve muitas variáveis, que o tornam um problema complexo a ser resolvido. Dentre estas variáveis pode-se citar [ALL90]:

- o ambiente no qual o sistema trabalha (ruidoso, silencioso, irregular);
- o tipo particular de microfone que capta o sinal da fala, incluindo considerações de direcionamento e posição;
- seleção do locutor, incluindo idade, sexo, língua de origem, sotaque, etc;
- velocidade da fala e entonação, que podem ocorrer para um mesmo locutor;
- variabilidade normal associada ao aparelho fonador, como o deslocamento do ar provocado pela língua, lábios, mandíbula, fossas nasais, etc.

Tentando limitar as possibilidades de variação, um projetista pode impor restrições à sua aplicação. Rich e Knight [RIC93] comentam cinco questões relacionadas ao projeto de sistemas de reconhecimento de fala que definem as restrições a uma aplicação:

- dependência do locutor versus independência do locutor - é difícil alcançar a independência do locutor devido às amplas variações de entonação e pronúncia. Por outro lado, um sistema dependente do locutor terá melhores resultados apenas quando o locutor que o treinou estiver utilizando;
- fala contínua versus isolada - o reconhecimento de palavras isoladas é mais fácil do que em discurso contínuo, pois os efeitos de borda fazem com que as palavras sejam pronunciadas diferentemente em contextos diferentes. Por exemplo, a palavra "horas" terá um som quando estiver na expressão "muitas horas" (som inicial de 'z') e outro quando estiver na expressão "marcar horas" (som inicial de 'r');
- tempo real versus processamento *off-line* - aplicações altamente interativas exigem que a resposta ocorra à medida em que as palavras sejam enunciadas, enquanto que outras aplicações permitiriam algum tempo de computação;

• vocabulário grande versus pequeno - reconhecer palavras que estejam confinadas a pequenos vocabulários é mais fácil do que trabalhar com grandes vocabulários;

• gramática ampla versus restrita - um exemplo de gramática restrita é aquele dos números de telefone: S ⇒ XXX-XXXX, onde X é qualquer número entre 0 e 9. As limitacões sintáticas e semânticas para uma gramática ampla são mais difíceis de representar, aumentando o espaço de busca para o reconhecimento.

Particularizando as vantagens citadas no tópico 3.2.3 para o problema de reconhecimento de fala, observa-se que as redes neurais podem oferecer grande auxilio na busca da solução. Características como a capacidade de generalização são importantes para um problema que possui como objeto de estudo algo que varia enormemente, como o sinal de fala.

3.3.2 Trabalhos correlatos

Vários trabalhos vêm sendo relatados sobre a utilização de redes neurais artificiais em tarefas de reconhecimento de fala.

Já comentado no tópico 2.4, o trabalho de Teuvo Kohonen destaca-se na literatura por ser um dos precursores e por ter atingido bons resultados.

No Brasil , um trabalho a ser destacado é o de Carrijo e Figueiredo [CAR92], onde, utilizando-se de quantização de vetores e rede multicamadas com algoritmo *backpropagation*, alcançaram uma taxa de acerto de 77,9% para um vocabulário de 40 palavras.

Outros modelos de redes neurais foram concebidos ou tornaram-se úteis para o reconhecimento de fala:

• *Recurrent*: 1987, Almeida e Pineda [MAR90];

• *Time-delay*: 1987, D.W. Tank e J.J. Hopfield [MAR90];

• *Adaptative resonance theory - ART*: Carpenter e Grossberg [WAN93].

3.4 Considerações Finais

Reconhecimento de padrões é um problema bastante estudado e abordado de diversas formas. Particularmente, o reconhecimento de fala mostra-se como uma área onde há ainda muito a ser explorado.

As redes neurais são uma técnica que auxiliam na tarefa de reconhecimento de padrões, dada suas características e vantagens frente à natureza não-estruturada dos padrões. Principalmente nos padrões de fala, onde uma estrutura exata é difícil de ser definida face a grande variabilidade existente, provocada por entonação, timbre, sotaque, etc. Visando diminuir esta variabilidade, algumas restrições são impostas às aplicações, procurando torná-las menos complexas.

Vários trabalhos vêm sendo desenvolvidos, demonstrando que há um futuro promissor na resolução do problema de reconhecimento de fala com a utilização de redes neurais.

4. INTERFACE DE RECONHECIMENTO DE VOZ PARA O SGCIF

4.1 Introdução

Esta dissertação originou-se do Trabalho de Conclusão de Curso intitulado "Construção de um protótipo de software comandado por voz" [HUG92], que se limitou a reconhecer dois comandos de voz.

A aplicação hipotética criada naquele trabalho compunha-se de um carro desenhado na tela do computador, que era movimentado a partir dos comandos "Esquerda" e "Direita" emitidos por um locutor e reconhecidos pelo protótipo.

O protótipo realizava a captação e digitalização do som através de um aparelho denominado de CI-500, que fornecia em torno de 7 Kbytes de dados para cada segundo de fala. Esta massa da dados era pré-processada visando extrair um vetor característico de 60 elementos binários, onde cada elemento representava a ocorrência ou não de uma oclusiva - modo de articulação do som onde o aparelho fonador impede a passagem de ar naquele momento - em cada um dos 60 instantes de tempo em que se dividia a palavra falada.

Utilizando-se de uma rede neural modelo Perceptron com treinamento supervisionado, alcançou-se uma taxa de acerto de 90% quando da utilização do protótipo pelo mesmo locutor que efetuou o treinamento. Porém esta taxa caía muito caso se aumentasse o número de palavras a serem reconhecidas.

A aplicação ora estudada - interface do Sistema Gerenciador de Central de Informação de Fretes (SGCIF) - envolve o reconhecimento de várias palavras, exigindo a construção de um novo protótipo que incorpore aprimoramentos ao protótipo descrito anteriormente [HUG92]. Nos próximos tópicos deste capítulo serão descritos os elementos que compõem esta nova aplicação e os estudos desenvolvidos para se atingir o reconhecimento de palavras da interface do SGCIF.

4.2 Sistema Gerenciador de Central de Informação de Fretes - SGCIF

Este sistema tem por finalidade auxiliar no gerenciamento de uma Central de Informação de Fretes (CIF). De acordo com o Departamento Nacional de Estradas de Rodagem (DNER) [apud PEZ93], as Centrais de Informação de Fretes são um serviço prestado, gratuitamente, pelo Departamento de Transportes e Terminais - DETER, aos transportadores e fornecedores de carga, funcionando como estrutura de apoio, informação e encaminhamento. Elas surgiram "da necessidade do controle dos gastos com combustíveis, da incerteza do transportador na obtenção da carga e na dificuldade das empresas em encontrar transporte para seus produtos, fazendo com que muito tempo, dinheiro e principalmente, combustíveis sejam gastos desnecessariamente" [LAP88].

De acordo com Pezzi [PEZ93], uma CIF funciona como difusora de informações a respeito do mercado de fretes. A troca de informações, entre transportadores e fornecedores de cargas através das CIFs, conduz a um ajuste mais adequado nas negociações dos fretes, além de promover a agilização e otimização dos transportes, evitando desperdícios com o deslocamento de veículos vazios.

As atividades principais envolvidas nas CIFs de Santa Catarina são [PEZ93]:

- cadastrar fornecedores e transportadores que desejem participar das Centrais;
- entrar em contato com fornecedores, quando estes não o fazem, a fim de estimular as ofertas de cargas;
- montar os formulários de "Bolsas de Cargas";
- atender os transportadores fornecendo informações sobre cargas de interesse dos mesmos e expedir os "Acordos de Cargas".

O Sistema Gerenciador de Central de Informação de Fretes (SGCIF) permite o cadastramento de informações como transportadores, fornecedores, veículos, cidades e cargas a serem transportadas. Baseado nestas informações, o sistema é capaz de definir rotas para os caminhões transportarem as mercadorias, visando uma maximização de lucros e minimização de custos.

O SGCIF foi desenvolvido na Universidade Federal de Santa Catarina pelo Grupo de Estudos em Transportes (GETRAN), sob coordenação dos professores Ricardo Miranda Barcia e Amir Mattar Valente. A versão 2.0, de Março de 1992, foi tomada como modelo para os estudos da interface por voz. Ela é executada em

microcomputadores padrão IBM-PC em sistema operacional DOS, possuindo uma interface baseada em caracter. O menu principal do SGCIF pode ser visualizado na figura 3.

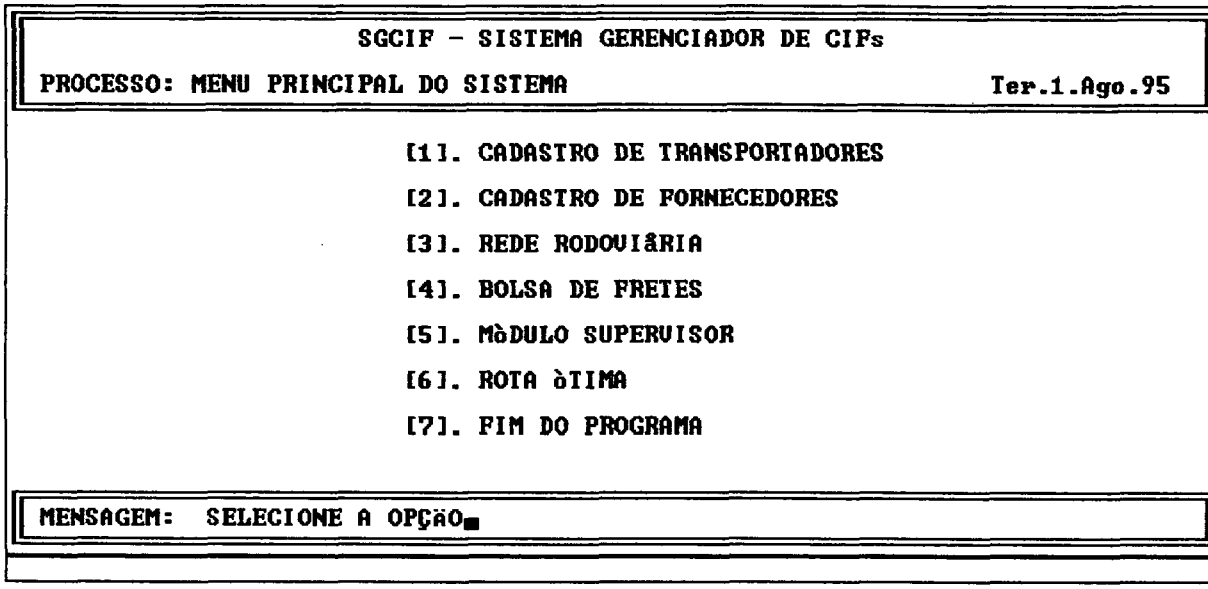

Figura 3: Menu principal do SGCIF

Visando melhorar a interação homem-máquina, além de oferecer uma interface de voz, resolveu-se migrar o SGCIF para um ambiente de interface gráfica, seguindo a tendência do mercado de informática em oferecer sistemas com interfaces gráficas mais amigáveis.

O ambiente escolhido foi o *Microsoft Windows*, pela significativa fatia de mercado que o mesmo ocupa e tende a ocupar nos próximos anos. Em relatório publicado pela *Redmond Group* [ANA92], esta empresa prevê que, até o ano de 1997, as duas versões do Windows (Windows NT e Windows/DOS) irão representar juntas 83% do total dos novos sistemas operacionais vendidos. Em pesquisa efetuada pela BYTE, 47% dos entrevistados afirmaram que Windows/DOS será o sistema dominante em 1997, enquanto que outros 17% atribuíram ao NT este papel [ANA92]. O protótipo do SGCIF em Windows foi desenvolvido utilizando-se do compilador Borland C++ Versão 4.0.

Em decorrência do ambiente Windows ser orientado a eventos, permitindo que um mesmo usuário possa acessar vários pontos da mesma aplicação "simultaneamente", a hierarquia de menus do SGCIF foi alterada procurando ser compatível com esta nova forma de interação. Assim, da hierarquia de menus existente no SGCIF para

DOS, apresentada no quadro 5, criou-se a nova hierarquia para o protótipo em Windows (quadro 6).

Menu Principal	Nível 1	Nível 2	Nível 3
Cadastro de Transportadores	Transportadores	1	
	Veículos	1	
	Relatórios	• Transportadores • Veículos • Retornar	
	Retornar		
Cadastro de Fornecedores	Fornecedores	1	
	Cargas	1	
	Relatórios	• Fornecedores • Cargas • Retornar	
	Retornar		
Rede Rodoviária	Nós da rede	1	
	Vizinhos	1	
	Relatórios	• Nós da rede • Vizinhos • Retornar	
	Retornar		
Bolsa de Fretes	Inclusão		
	Exclusão		
	Alteração		
	Consulta		
	Relatórios	Todos	
		com Restrições	• Cidade Destino • Características do veículo • Tipo de carga • Valor do Frete • Prazo de Entrega • Distância a percorrer
		Retornar	
	Retornar		
Módulo Supervisor	Relatório Movimentação dos Transportadores	2	
	Relatório Movimentação dos Fornecedores	2	
	Relatório Movimentação dos Fretes	2	

	Registrados		
	Emissão da tabela das medidas de desempenho		
	Retornar		
Rota Ótima	Busca da rota		
	Parâmetros		
	Retornar		
Fim do programa			

Quadro 5: Hierarquia de menus do SGCIF para DOS.

Nota: 1 = Opções Inclusão, Exclusão, Alteração, Consulta e Retorna

2 = Opções De um período, Na totalidade e Retorna

Menu Principal	Nível 1	Nível 2	Nível 3
Cadastros	Transportadores		
	Veículos		
	Fornecedores		
	Cargas		
	Rede Rodoviária	• Nós • Vizinhos	
	Fretes		
	Terminar		
Relatórios	Transportadores		
	Veículos		
	Fornecedores		
	Cargas		
	Rede Rodoviária	• Nós • Vizinhos	
	Fretes	• Todos • com Restrição • Formulário	
	Supervisão	Movimentação	• Transportadores • Fornecedores • Fretes
		Desempenho	
Rota Ótima	Busca da rota		
	Parâmetros		
Terminar			

Ajuda	Conteúdo		
	Localizar		
	Como usar Ajuda		
	Sobre		

Quadro 6: Hierarquia de menus do SGCIF para Windows.

Conforme as atuais restrições para o reconhecimento de fala - tratadas no capítulo 3 - é necessário limitar o vocabulário a ser reconhecido. Scapin [SCA86] afirma que se a entrada por voz se alterna com outros modos de entrada, é útil distinguir entre dados e comandos, utilizando a voz para os comandos e, por exemplo, teclado para dados. Assim, o reconhecimento de fala neste trabalho fica limitado às palavras do quadro 6, visto que estas estão relacionadas aos comandos que operam o protótipo do SGCIF em Windows.

4.3 Fases do reconhecimento de fala

A tarefa de reconhecer palavras faladas pode ser divida em três fases: Busca de Sinais, Pré-processamento e Processamento. Cada qual realiza transformações nos dados recebidos, como pode ser visto na figura 4.

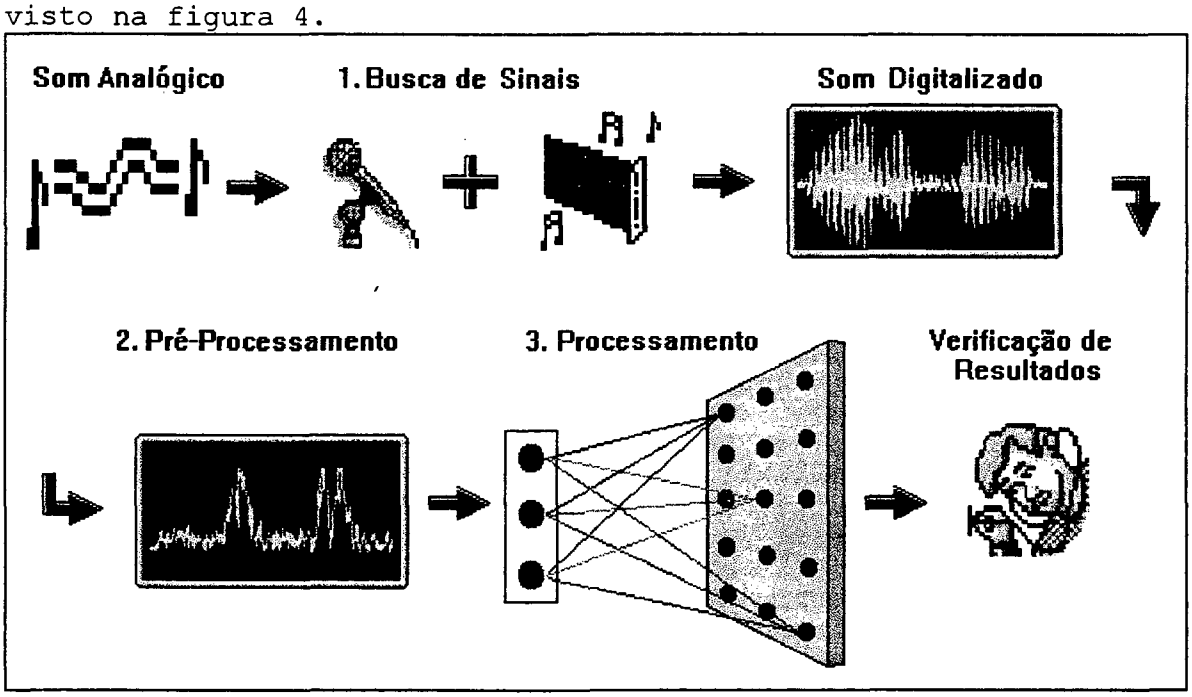

Figura 4: As fases empregadas para reconhecimento de fala.

4.3.1 Busca dos sinais

A primeira fase, denominada de Busca dos Sinais ou Captação, realiza a captação do som da fala e sua transformação em sinal digital para manipulação por um computador digital.

"Na sua forma natural os sons são análogos. Eles aumentam e diminuem, alterando a freqüência e amplitude em suaves progressões. Obviamente, o computador não pode armazenar o som como ele é. Para armazenar o som, o formato da informação precisa ser alterado para que o computador possa armazenar ou manipular os sons na memória. Em outras palavras, a informação precisa ser digitalizada. Ela precisa ser convertida em números digitais com os quais o computador pode trabalhar" [MUN94].

O som é captado através de um microfone, o qual está acoplado a uma placa digitalizadora. Placas de som, tais como a *Sound Blaster* ou *Fusion CD 16*, realizam operações de digitalização de sinais sonoros. Detalhes sobre o som e sua digitalização, podem ser consultados em Matras [MAT91] e Stolz [STO93] respectivamente.

4.3.2 Pré-processamento

O pré-processamento de sinais é uma forma de organizar ou ajustar os sinais captados procurando torná-los passíveis de processamento na fase posterior do processo de reconhecimento de fala. Ele é responsável por gerar o vetor característico do padrão a ser analisado (tópico 3.2.2).

Esta fase pode ser subdividida no que Bezdek e Pal [BEZ92] chamam de:

- pré-processamento: utilização de técnicas de "limpeza" do sinal, como normalização, escala, suavização, etc;
- extração de características: eliminação de sinais redundantes ou insignificantes através de seleção ou transformação.

4.3.3 Processamento

Esta fase é responsável por realizar o reconhecimento do padrão, normalmente por agrupamento ou classificação [BEZ92]. A diferença entre eles, é que o agrupamento utiliza-se freqüentemente

de técnicas de aprendizado não-supervisionado para criar grupos de padrões dentro do universo a ser reconhecido, enquanto que a classificação freqüentemente utiliza-se de técnicas de aprendizado supervisionado para rotular , se for o caso, todos os padrões do universo a ser reconhecido.

Uma vez realizado o processamento, pode-se tomar a atitude correspondente ao padrão reconhecido, de acordo com a aplicação.

4.4 Desenvolvimento do trabalho

Para sair do reconhecimento de dois comandos [HUG92] e alcançar o objetivo de operar por voz os menus do protótipo do SGCIF em Windows, o trabalho de pesquisa foi dividido em estágios.

Em todos os estágios, as fases empregadas para realizar a tarefa de reconhecimento de fala foram as mesmas. Porém, cada estágio definido procurava aprimorar uma das fases, a saber:

1. fase de Processamento: utilizar outro modelo de rede neural;
2. fase de Pré-processamento: procurar outros modos de pré-processamento;
3. fase de Busca de Sinais: migrar a captação do som do equipamento CI-500 para uma placa de som comum em microcomputadores PC, como por exemplo a *Sound Blaster*.

Os estágios foram desenvolvidos na ordem inversa da execução das fases, para que pudesse ocorrer uma evolução segura do trabalho e a certeza de que os novos resultados alcançados eram decorrentes apenas das modificações implementadas naquele momento.

4.4.1 Estágio 1 - A utilização do modelo de Kohonen

O objetivo do primeiro estágio era de substituir o modelo de rede neural empregado. Na aplicação do protótipo construído em [HUG92], o modelo Perceptron foi utilizado para o reconhecimento das palavras "Esquerda" e "Direita". Esta mesma aplicação serviu de base para o estágio 1 deste trabalho.

O modelo de rede neural utilizado para a tarefa do reconhecimento de fala neste trabalho foi o de Kohonen, também conhecido por Mapas Auto-organizáveis de Kohonen. Optou-se por trabalhar com Redes de Kohonen por alguns motivos, dentre os quais:

- a própria audição humana cria estruturas auto-organizáveis durante a fase de aprendizagem [KOH90];
- as Redes de Kohonen foram originalmente concebidas para trabalhar em aplicações de reconhecimento de fala [KOH90] [LIP87];
- a relação entre os benefícios dos potenciais desta rede em comparação ao tempo dispendido para treinamento é bom, se comparado com outros modelos, como por exemplo *backpropagation* [CAR88].

Huntsberger e Pongsak [HUN92] explicam que os modelos de redes neurais para reconhecimento de padrões podem ser especificados em dois tipos: os supervisionados (como Hopfield e Perceptron) e os não-supervisionados (como Kohonen e Carpenter/Grossberg). Tipos de redes como Perceptron e Hopfield sofrem de problemas de convergência a mínimos locais e de instabilidade de memória. Parte destes problemas podem ser atribuídos ao uso da regra Delta para a atualização dos pesos. Sistemas como Carpenter/Grossberg e Kohonen são baseados em estudos biológicos da organização e dinâmica da memória. Alguns trabalhos demonstram que no modelo de Carpenter/Grossberg, o comportamento replicável de um sistema parece ser sensível aos valores exatos dos parâmetros de inicialização do sistema. Já no modelo de Kohonen não há este problema, além de utilizar uma regra linear para atualização dos pesos, o que o torna um modelo computacionalmente atrativo.

A arquitetura proposta por Kohonen consiste de uma rede neural plana (bidimensional) formando uma camada competitiva de neurônios. Cada neurônio desta camada está amplamente conectado à camada de entrada através de pesos sinápticos, que são ajustados por um processo de aprendizado não-supervisionado. Graficamente, pode-se visualizar na figura 5.

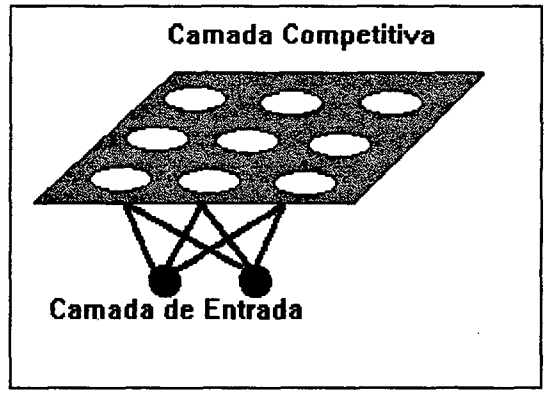

Figura 5: Rede de Kohonen

A camada bidimensional é conhecida por competitiva devido ao fato de que os neurônios, durante o processo de treinamento, competem pelo direito de aprender (ou ajustar-se) a reconhecer um determinado padrão de entrada que é apresentado.

Uma das características importantes dos Mapas (ou Redes) de Kohonen é que as localizações das respostas tendem a se ordenar como se um sistema significativo de coordenadas aos diferentes padrões de entrada fosse criado sobre a rede.

O algoritmo de aprendizagem não-supervisionada realiza ajustes nos pesos sinápticos das conexões entre a camada de entrada e a competitiva, visando a auto-organização dos neurônios pela alteração dos pesos do neurônio vencedor e de seus vizinhos, tornando esta região mais sensível ao padrão do estímulo de entrada.

Um neurônio da camada competitiva é considerado vencedor quando o vetor do padrão de entrada possui a mínima distância Euclidiana com o vetor de pesos correspondente, isto é, a diferença entre os valores do vetor de entrada e os do vetor de pesos é a menor em relação às diferenças dos outros neurônios.

O algoritmo de auto-organização é basicamente o seguinte [LIP87]:

1. Inicialização dos pesos :

Antes de qualquer operação na rede, inicializar todos os pesos das sinapses com valores aleatórios pequenos.

2. Apresentar uma nova entrada.

3. Calcular as distâncias entre os neurônios da camada competitiva e os neurônios da camada de entrada:

Para cada neurônio da camada competitiva, calcular a distância dj entre o neurônio j da camada competitiva e cada neurônio i da camada de entrada.

$$dj = \Sigma (xi(t) - wij(t))^2$$

onde : $xi(t) \Rightarrow$ valor do neurônio de entrada i no tempo t.

$wij(t) \Rightarrow$ peso sináptico entre o neurônio de entrada i e o neurônio de saída j no tempo t.

4. Selecionar o neurônio vencedor

Seleção do neurônio j cuja distância dj seja a menor possível.

5. Atualização de pesos da vizinhança do neurônio vencedor

A atualização de pesos é realizada no próprio neurônio vencedor bem como numa área que abrange NE(t) neurônios à sua volta e essa atualização obedece o seguinte cálculo :

$wij(t+1) = wij(t) + \eta(t).(xi(t) - wij(t))$

onde : $\eta(t)$ $\Rightarrow$ taxa de aprendizagem que deve ser diminuida no decorrer do treinamento.

NE (t) $\Rightarrow$ número de neurônios vizinhos afetados pela atualização de pesos.

6. Voltar ao passo 2 para novo treinamento.

Quadro 7: O algoritmo da rede de Kohonen.

Tanto NE(t) quanto $\eta(t)$ devem decrescer ao longo do tempo, para que o processo de convergência, acelerado no princípio, torne-se mais suave com o passar do tempo, evitando oscilações indesejáveis.

Se o mapa auto-organizável for utilizado como classificador de padrões, onde as respostas são agrupadas em sub-conjuntos, cada um correspondendo a uma classe discreta de padrões, então o problema é considerado um processo de decisão e deve ser tratado de modo diferente. Certas áreas identificadoras dos padrões de entrada podem, após o término do aprendizado, estarem com limites muito próximos, coincidentes, ou até sobrepostos. Para resolver tal problema, ou seja, melhor delinear os limites das áreas, Kohonen apresentou três métodos, que chamou de LVQ1, LVQ2 e LVQ3 [KOH90].

Os métodos de quantização do vetor de aprendizado (*LVQ - Learning Vector Quantization*) permitem que, através de processos iterativos, sejam reforçados, ou inibidos, os pesos de determinados neurônios da rede, caso eles proporcionem uma resposta correta, ou incorreta, respectivamente, a um padrão de entrada.

No presente estudo apresentar-se-á apenas o LVQ1, que foi o método utilizado, pelo qual os pesos são alterados da seguinte forma [KOH90]:

* $wj(t+1) = wj(t) + \eta(t)(x(t) - wj(t))$, caso x tenha sido classificado corretamente;
* $wj(t+1) = wj(t) - \eta(t)(x(t) - wj(t))$, caso x tenha sido classificado incorretamente;
* $wk(t+1) = wk(t)$, para k diferente de j

onde: j = índice do neurônio vencedor, e

$\eta(t)$ = taxa de aprendizado, que é decrescida monotonicamente ao longo do tempo. Como este é um processo de sintonia fina, recomenda-se utilizar valores para $\eta(0)$ entre 0,01 e 0,02.

Quadro 8: O algoritmo de LVQ1.

Desta forma, então, procura-se eliminar distorções na interpretação espacial da rede, criando regiões bem definidas que melhor classificam os padrões de entrada.

Utilizando-se estes algoritmos, foi realizada uma primeira tentativa, procurando reconhecer cinco comandos que movimentassem o carro desenhado na tela do computador. A rede montada para este teste era composta por 41 neurônios de entrada, representando o vetor característico do padrão, o qual é formado por 40 instantes de tempo e o tamanho do vetor de som digitalizado, e composta por 49 (7x7) neurônios na camada competitiva.

Os comandos e os resultados obtidos foram os seguintes:

Comando	Taxa de Acerto (%)
Esquerda	94
Direita	72
Ande	70
Pare	64
Fim	72
Média Geral do Reconhecedor	**74,40**

Quadro 9: Taxas de acerto da primeira tentativa

4.4.2 Estágio 2 - Outras formas de pré-processamento

Com a utilização da Rede de Kohonen, houve a facilidade de se utilizar números reais como valor para cada neurônio de entrada, ao invés dos valores binários utilizados até então.

A forma de pré-processamento adotada calculava as médias em 60 instantes de tempo, porém verificava se em cada instante havia a ocorrência de uma oclusiva, limitando os valores a 0 ou 1 (tópico 4.1).

A partir deste estágio, começou-se a utilizar a diferença entre a média superior e a média inferior em cada instante, que representava a amplitude do sinal naquele instante. Desta forma, há a preservação das principais características da forma de onda do som, ocorrendo uma considerável diminuição no tamanho do vetor característico, como pode ser observado na figura 6.

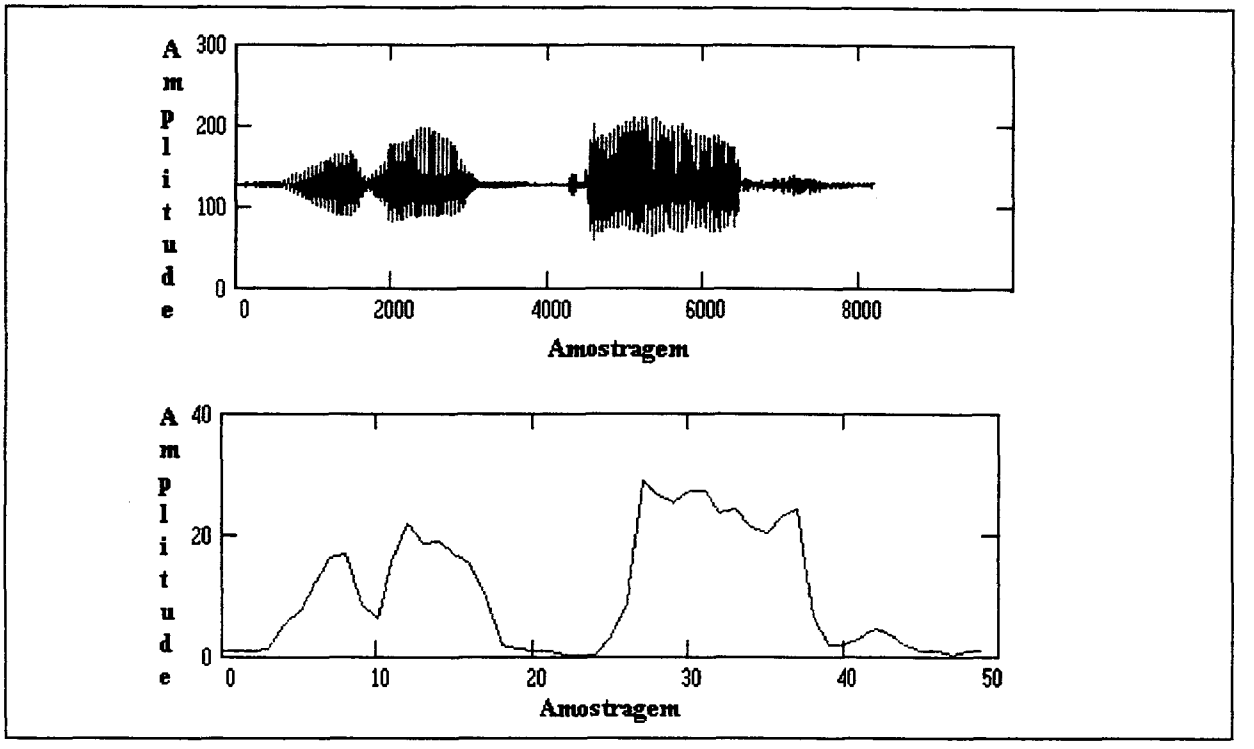

Figura 6: Gráfico da forma da onda e do vetor característico da palavra Relatórios.

Com esta nova forma de pré-processamento, realizou-se uma segunda tentativa, procurando reconhecer sete comandos com uma taxa de acerto maior. Os sete comandos foram escolhidos aleatoriamente dentro do conjunto de palavras que eram encontradas no SGCIF Versão 2.0.

O pré-processamento envolveu a extração de médias em 40 instantes de tempo; cálculo da diferença entre as duas médias de um mesmo instante de tempo; transformação desta diferença em valores para os neurônios de entrada, sendo 40 neurônios representando os intervalos de tempo, 21 representando reforços de alguns intervalos e 1 neurônio representando o tempo de fala captado para este padrão, totalizando 62 neurônios na camada de entrada.

A camada competitiva constava de 49 neurônios (7x7), sendo que os pesos foram inicializados com a média geral dos valores de todos os padrões do conjunto de treinamento, diminuída de um valor aleatório pequeno, calculado para cada sinapse entre a camada de entrada e a de competição.

Após o ajuste fino, as taxas de acerto de cada palavra foram modificadas conforme o quadro 10. Para a verificação da taxa de

acerto, foram realizados testes com 50 exemplos de cada comando a ser reconhecido.

Comando	Taxa de Acerto (%)	
	Treino Normal	LVQ1
Cadastros	94%	98%
Lançamentos	88%	96%
Cálculo	84%	86%
Emissão	80%	84%
Manutenção	92%	92%
Outros	74%	76%
Fim	92%	92%
Média Geral do Reconhecedor	**86,29%**	**89,42%**

Quadro 10: Taxas de acerto da segunda tentativa

Os resultados alcançados foram mais animadores em relação ao primeiro estágio, principalmente se for considerado que, com o aumento do número de palavras, era esperado uma queda na taxa de acerto. Apesar disto, verificou-se a possibilidade de se utilizar a Transformada Rápida de Fourier (*FFT - Fast Fourier Transform*), por ser uma forma de pré-processamento muito indicada na literatura [FU82] [ALL90] [CAR92].

4.4.3 A FFT como Ferramenta para o Pré-processamento

No início do século XVIII, o matemático francês Jean Baptiste Fourier realizou pesquisas para descobrir como o som poderia ser dividido em elementos basicamente idênticos. Para isso, ele desenvolveu um processo matemático que divide um som em um número finito de ondas senóides, o qual é atualmente conhecido como Análise de Fourier. Esse processo também pode ser utilizado em ordem contrária, de modo a produzir sons específicos artificialmente (Síntese de Fourier) [BRU95].

A Transformada Rápida de Fourier é uma forma de processamento de sinais largamente aplicada. A FFT pode ser utilizada por exemplo em aplicações de comunicação, radares, sonares, processamento de sinais, processamento da fala, área de engenharia biomédica, simulações, síntese musical, entre outros [TUC92].

Com a FFT tem-se condições de identificar ou distinguir as senóides de diferentes freqüências e suas respectivas amplitudes que são combinadas para formar a onda sonora.

Com a utilização da FFT pode-se ajustar falhas ocorridas na busca (digitalização) dos sinais analógicos, bem como diminuir a quantidade de sinais existentes. Dessa forma, é possível aplicar uma técnica mais precisa que ao mesmo tempo possibilita essa diminuição no número de elementos de processamento e também pode corrigir possíveis distorções.

- Como a Transformada Rápida de Fourier funciona

Uma função no domínio do tempo é traduzida pela FFT em uma função no domínio da freqüência, onde a função pode ser analisada pelas freqüências que nela são encontradas [COD92].

A essência da FFT de uma onda é a decomposição ou separação da onda em uma soma de senóides de freqüências diferentes. A representação gráfica da FFT é um diagrama que mostra a amplitude e a freqüência de cada uma das diferentes senóides.

Mesmo que o número de freqüências diferentes existentes no som resulte em uma grande quantidade de dados que precisam ser processados, é válido utilizar-se desse processo matemático para otimizar os sinais obtidos porque existe uma diminuição considerável na quantidade de sinais existentes sem que exista perda das características principais da onda.

No uso da FFT, devem ser tomados certos cuidados. A matriz utilizada como entrada para a função deve ser composta por 2^n elementos, onde n deve ser inteiro e maior que 2. Como resultado a FFT retorna uma outra matriz contendo $(2^{n-1} + 1)$ elementos. Por exemplo, quando se utiliza como entrada um vetor de 256 elementos, obtem-se um outro vetor com 129 elementos. A figura 7 demonstra uma aplicação da FFT sobre a forma de onda da palavra "Relatórios" (figura 5).

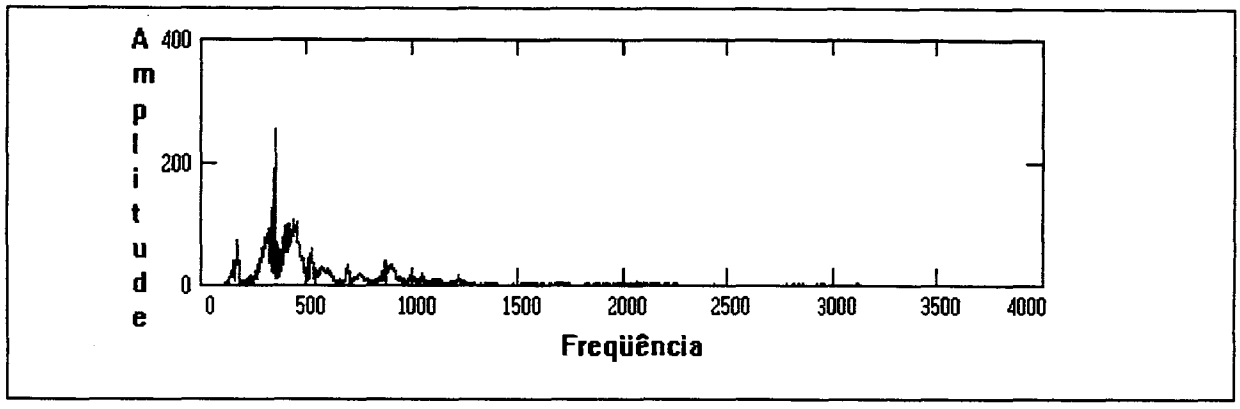

Figura 7: Gráfico da FFT da onda da palavra Relatórios.

4.4.4 Testes finais do segundo estágio

Para a realização dos testes relativos ao reconhecimento de palavras através da fala optou-se por utilizar como objeto dos testes somente quatro palavras: Documentos, Dólar, Estrangeiras e Voltar [BRU95]. Estas quatro palavras foram escolhidas por fazerem parte de um menu da BBS (*Bulletin Board System*) da FURB, sem haver preocupação com suas formas fonéticas. O conjunto de palavras foi escolhido somente para realizar o comparativo entre as duas formas de pré-processamento, não havendo assim a necessidade de ser coincidente com os conjuntos utilizados em testes anteriores.

Cada uma das quatro palavras utilizadas nos testes foram faladas cinco vezes, sendo captadas a uma taxa amostral em torno de 11 Khz, em som mono de 8 bits. Os vinte padrões obtidos (conjunto de treinamento) foram gravados sem o pré-processamento, de forma que esses mesmos padrões pudessem ser utilizados tanto no pré-processamento através de médias como no pré-processamento através da FFT. Assim procura-se garantir que os resultados obtidos para padrões idênticos de entrada e processamento na rede neural tenham sua diferenciação apenas na etapa de pré-processamento.

Para a verificação das saidas obtidas e da eficiência do reconhecimento foram faladas 100 vezes cada uma das 4 palavras a reconhecer. Foram apresentadas as 400 palavras à rede neural com o objetivo de identificar a quantidade de ocorrências para cada uma das saídas da rede neural. Essas palavras foram previamente gravadas e para cada tipo de pré-processamento foram utilizadas as mesmas 400 palavras. Com isso foram obtidos os resultados considerando-se os mesmos padrões de entrada.

• O teste com o pré-processamento através de médias

Primeiramente foi realizada a experiência com o pré-processamento através das médias. Para tanto, os padrões até então gravados na forma original da onda sonora digital foram pré-processados através do cálculo de médias. Logo após, com base nesses padrões pré-processados, foi realizado o treinamento da rede neural e também o ajuste fino dos pesos.

A transformação proporcionada pelo pré-processamento através de médias gerou um vetor característico de 60 elementos, cada um representando a diferença entre a média superior e a inferior em um instante de tempo.

Os resultados obtidos podem ser visualizados no quadro 11. Pode-se considerar que as taxas de erro encontradas para o pré-processamento dos sinais componentes de uma onda sonora, através de médias, são bastante aceitáveis devido à simplicidade desse tipo de pré-processamento.

Comando	Taxa de Acerto(%)	
	Médias	FFT
Documentos	93	66
Dólar	91	62
Estrangeiras	89	63
Voltar	74	58
Média Geral do Reconhecedor	86,75	62,25

Quadro 11: Comparativo entre taxas de acerto utilizando pré-processamento através de médias e de FFT.

• O Teste Com o Pré-processamento Através da FFT

Após realizado o pré-processamento dos padrões originais através da FFT foram realizados o treinamento da rede neural e o ajuste fino de pesos. As mesmas 400 palavras faladas foram colocadas em prova para testar a eficiência do reconhecimento na aplicação dessa forma de pré-processamento.

Cada palavra falada era armazenada em um vetor de 8192 elementos. O vetor resultante da FFT possuía 4093 elementos, sendo ainda um vetor de grande dimensão para ser processado. Desta forma se fazia necessário diminuí-lo a um vetor de dimensões adequadas,

onde foi aplicado o cálculo de médias, similar ao exposto, procurando reduzir o vetor a 60 elementos.

Os resultados obtidos com o pré-processamento de padrões através da FFT, que podem ser vistos no quadro 11, mostram altas taxas de erro encontradas no reconhecimento das 400 palavras apresentadas à rede neural.

Após os testes realizados com os dois tipos de pré-processamento, preocupando-se em utilizar os mesmos padrões com os sons originais e a mesma forma de processamento através da rede neural, foi possivel optar por uma decisão bastante clara.

O pré-processamento através da FFT é considerável mas não supera, da forma como foi estudado, o pré-processamento através de médias. Acredita-se que isto se deve ao fato de se estar utilizando a abordagem global para o reconhecimento (tópico 2.3.2). Ao se utilizar a FFT ocorre a perda da noção de tempo, o que influencia o reconhecimento nesta abordagem. Exemplificando de forma tosca, pronunciar a palavra "Roma" é diferente de pronunciar a palavra "amor" devido ao tempo em que cada letra é pronunciada. Porém, ao se realizar a FFT, as duas palavras possuiriam vetores representativos semelhantes. Isto pode ser observado nos gráficos da figura 8. O gráfico f_j representa a FFT do vetor F e o gráfico g_j representa a FFT do vetor G, sendo:

F = [0,0,1,2,3,4,5,6,7,8,9,10,0,0,0,0]

G = [0,0,0,0,10,9,8,7,6,5,4,3,2,1,0,0]

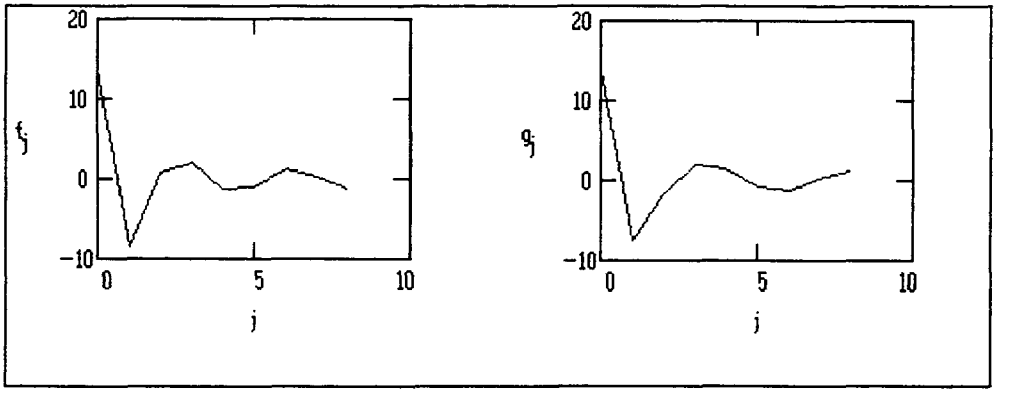

Figura 8: Gráfico da FFT dos vetores F e G.

Além de apresentar melhores resultados, o pré-processamento através de médias é computacionalmente mais "leve", sendo executado de forma mais rápida. Segundo [SCA86], um tempo de resposta aceitável é aquele de até 2 segundos. Desta forma, após o usuário ter dado o comando por voz deveria ocorrer a ação correspondente

dentro de 2 segundos. Assim, para que isto ocorra, deve haver uma preocupação com a complexidade dos algoritmos utilizados para o reconhecimento.

4.4.5 Estágio 3 - Busca de Sinais

Em virtude do protótipo do SGCIF ser executado em ambiente Windows, se fazia necessário realizar a captação e digitalização dos sinais utilizando um dispositivo que já fosse manipulado por este ambiente. Além disto, a placa CI-500 não é um equipamento conhecido nem utilizado em escala comercial, o que dificultaria a utilização do reconhecimento de fala através deste protótipo em vários locais e em várias configurações de computadores.

Uma das grandes vantagens do ambiente Windows é a de oferecer ao programador de aplicações uma independência de hardware, ou seja, o programador não precisa se preocupar com os detalhes de configuração de hardware, mas apenas utilizar-se das funções disponíveis nas APIs (*Application Programming Interface*) do Windows. Estas APIs realizam a compatibilização entre a aplicação e o hardware instalado, e são consideradas como interfaces de baixo nível por se tratarem de um conjunto de funções acessíveis por linguagens de programação como C e Pascal. Desta forma, muitas APIs possuem interfaces de mais alto nível, visando abstrair uma série de detalhes e facilitar a programação. No caso de tratamento de dispositivos de multimídia, entre eles o de som, o Windows possui a interface de mais alto nível conhecida como MCI (*Media Control Interface*).

A interface de controle de mídia - MCI é uma linguagem de comandos padronizada para comunicação com dispositivos de multimídia, como *compact disc*, áudio em forma de onda (*wave*) e MIDI (*Musical Instrument Digital* Interface), videodiscos, arquivos AVI (*Audio-Video Interleaved*), mixer de áudio, etc. A especificação MCI, liberada em 1991 pela Microsoft e IBM, define um conjunto de comandos básicos, que podem ser aplicados a qualquer dispositivo em geral, e comandos estendidos para tipos específicos de dispositivos [MUS93]. Estes comandos estão disponíveis através de dois tipos de interface de programação dentro da MCI [MIC94]:

- a interface de strings de comando: permite que o programador use comandos na língua inglesa para se comunicar com os dispositivos MCI. Por exemplo, a string de comando seguinte reproduz um arquivo WAVE chamado "TIMPANI.WAV": ***play timpani.wav***. A interface de strings de comando foi projetada para ser usada com programação de alto nível e ambientes de autoria, como o *Microsoft*

Visual Basic e *Asymetrix ToolBook*. As aplicações provêem uma interface baseada em texto permitindo aos usuários controlar os dispositivos MCI através do uso da interface de strings de comando;

• a interface de mensagens de comando: utiliza o paradigma da passagem de mensagens para se comunicar com os dispositivos MCI. Por exemplo, o fragmento seguinte de código realiza a mesma operação do exemplo anterior de string de comando:

```
mciSendCommand(wDeviceID,                    /* device ID */
               MCI_PLAY,                     /* command message */
               0,                            /* flags */
          (DWORD)(LPVOID) &mciPlayParms); /* parameter block */.
```

A interface de mensagens de comando foi projetada para ser utilizada por aplicações que requeiram uma interface em linguagem C para controlar os dispositivos de multimídia.

Porém, como a MCI procura abstrair vários detalhes de implementação, acaba por conseqüência não permitindo um controle muito próximo dos dispositivos de multimídia. Desta forma, para se alcançar o controle do dispositivo de som que realiza as operações de digitalização e sintetização de forma de onda (WAVE), necessário para a captação e tratamento da fala, foi preciso mesclar a utilização de alguns comandos da MCI, onde era possível facilitar a implementação, com as funções de baixo nível da API do Windows.

4.5 Resultados alcançados

Cumpridos os estágios definidos, passou-se a treinar a rede neural para o reconhecimento das opções de menu do SGCIF. Estas opções foram divididas em 7 grupos visando facilitar o reconhecimento através da rede neural. A divisão foi possível em razão de apenas poucas opções precisarem estar disponíveis simultaneamente em dado momento de operação do sistema.

A rede neural construída possui 81 (9x9) neurônios na camada competitiva e 50 neurônios na camada de entrada. A camada de entrada é preenchida por um vetor característico, extraído pelo pré-processamento através das médias, realizado sobre o sinal da fala captado com uma taxa amostral de 8 Khz, em som mono de 8 bits, com o microfone a cerca de 10 cm do locutor.

Cada um dos sete grupos possui sua própria matriz de pesos. O protótipo utiliza a matriz correspondente ao grupo de acordo com o contexto onde o usuário se encontra naquele instante de operação do sistema. Para cada grupo foi realizado um treinamento não-supervisionado da rede neural, contando sempre com seis exemplares de cada palavra. O treinamento foi executado em 20.000 iterações. O ajuste fino ocorreu em 10.000 iterações, utilizando os seis exemplares do primeiro treinamento somados a mais quatorze exemplares de cada palavra falada pertencente ao grupo.

Os grupos e sua respectiva taxa de acerto podem ser visualizados nos quadros 12 a 18. As estatísticas estão baseadas em 50 amostras de cada palavra em cada grupo.

Comando	Taxa de Acerto (%)
Cadastros	96
Relatórios	82
Rota Ótima	60
Terminar	90
Ajuda	64
Média Geral do grupo 1	**78,4**

Quadro 12: Taxas de acerto do Grupo 1

Comando	Taxa de Acerto (%)
Transportadores	92
Veículos	76
Fornecedores	92
Cargas	82
Rede Rodoviária	96
Terminar	76
Fretes	74
Supervisão	58
Média Geral do grupo 2	**80,75**

Quadro 13: Taxas de acerto do Grupo 2

Comando	Taxa de Acerto (%)
Busca da Rota	98
Parâmetros	96
Média Geral do grupo 3	**97**

Quadro 14: Taxas de acerto do Grupo 3

Comando	Taxa de Acerto (%)
Conteúdo	94
Localizar	70
Como usar	72
Sobre	92
Média Geral do grupo 4	**82**

Quadro 15: Taxas de acerto do Grupo 4

Comando	Taxa de Acerto (%)
Nós	96
Vizinhos	98
Média Geral do grupo 5	**97**

Quadro 16: Taxas de acerto do Grupo 5

Comando	Taxa de Acerto (%)
Todos	82
com Restrições	86
Formulários	88
Média Geral do grupo 6	**85,33**

Quadro 17: Taxas de acerto do Grupo 6

Comando	Taxa de Acerto (%)
Movimentação	98
Desempenho	98
Média Geral do grupo 7	**98**

Quadro 18: Taxas de acerto do Grupo 7

A média geral do reconhecedor, tomando como base de cálculo as médias de cada palavra em seu grupo, ficou em 84,84%.

4.6 Considerações Finais sobre a Interface

A divisão do problema de reconhecimento de fala em três fases principais estabeleceu a possibilidade de aprimorar os estudos em cada uma delas individualmente, facilitando o desenvolvimento deste projeto. A relativa independência existente entre as fases

assegurou que os avanços realizados não causariam danos externos à fase.

É fundamental a importância dos métodos de pré-processamento no atual estado da computação, onde ainda é necessário reduzir ou comprimir os dados (voz), para que haja um tratamento em um tempo considerado satisfatório. Nesta compressão de dados é imprescindível evitar a perda das características essenciais da informação. Mesmo que no futuro seja resolvido o problema do manuseio de grandes volumes de dados, ainda poderá ser interessante efetuar uma alteração ou melhora em certos tipos de dados para um melhor processamento e até mesmo um aumento de performance.

O modelo de rede neural adotado - Kohonen - mostrou-se satisfatório para o processamento, utilizando-se da abordagem global para reconhecimento de fala, alcançando uma taxa de acerto geral de 84,84%. Este valor demonstra que as pesquisas realizadas no desenvolvimento deste trabalho atingiram os objetivos, se comparado com resultados obtidos em outros trabalhos similares, como [CAR92] - 77,9% e [KOH90] - 92%.

5. CONCLUSÃO

5.1 Conclusões sobre o trabalho

Várias formas de interface natural foram comentadas ao longo deste trabalho, induzindo a imaginar o modo facilitado de interação homem-máquina que irá existir até o final desta década.

A fala irá certamente abrir muito mais as portas do mundo da computação e da tecnologia aos leigos, facilitando o acesso e manuseio de vários tipos de equipamentos computacionais. Trabalhos como este, por exemplo, demonstram estas facilidades, mostrando a viabilidade e potencialidade destes tipos de interface.

Seguindo as restrições impostas pela tecnologia disponivel, o protótipo desenvolvido alcançou bons resultados, comprovando que certos tipos de aplicações já podem se utilizar destes dispositivos de entrada para facilitar a operação. Aplicações mais complexas necessitam não somente de um aperfeiçoamento na tarefa de reconhecimento de voz, mas também de uma combinação de outros meios de entrada/saída.

O processamento para reconhecimento dos padrões de fala realizado através do modelo de rede neural Kohonen, demonstrou ser capaz de atender às exigências de um trabalho como a implementação de interface para o SGCIF. A rede, através de um treinamento principalmente não-supervisionado, conseguiu assimilar os diferentes padrões alcançando uma taxa de acerto considerada boa: 84,84%.

O desenvolvimento do trabalho através dos seus estágios permitiu uma evolução segura, garantindo que os resultados alcançados em cada estágio auxiliaram, e não interferiram, na busca da solução.

O objetivo de controlar a operação dos menus do SGCIF através da voz foi atingido, demonstrando que a utilização de redes neurais artificiais é viável em aplicações desta natureza.

5.2 Limitações e Sugestões Futuras

Protótipos deste porte demonstram satisfatoriamente que há possibilidade de se utilizar comandos vocais para a operação de softwares. Contudo, interfaces naturais necessitam de sistemas capazes de compreender a fala contínua. Para isto são necessários avanços tecnológicos significativos que certamente passam por uma abordagem multidisciplinar, envolvendo aspectos acústicos, fonéticos, fonológicos, léxicos, sintáticos, semânticos e pragmáticos. A união destas áreas é que deve ser estudada em futuros trabalhos, visando ultrapassar as atuais barreiras encontradas.

Por outro lado, não se pode esquecer os aspectos ergonômicos oriundos deste novo tipo de interface. Pois se de um lado há uma melhora na forma como ocorre a atual interação entre o homem e a máquina, por outro originará novos problemas não imaginados até então. Scapin [SCA86] lembra que cada nova tecnologia faz surgir novos problemas em relação à tecnologia precedente. Portanto, outra área que deve ser arrolada em novos estudos é a da ergonomia de software, principalmente frente às novas interfaces vislumbradas para um futuro próximo.

Procurando vencer as restrições impostas por limitações de capacidade, a utilização de hardware específico para reconhecimento de padrões pode ser de grande auxílio. *Digital Signal Processor* (DSP) é um tipo de circuito que vem sendo utilizado por alguns pesquisadores a fim de implementar determinados algoritmos de tratamento da fala, tornando o processamento total do sinal de fala bem mais rápido [MEI93]. Além deste tipo, redes neurais implementadas em hardware poderiam realizar o processamento de forma quase instantânea, diminuindo ainda mais o tempo total de processamento.

Em relação ao protótipo desenvolvido, a tarefa de reconhecimento pode ainda ser realizada pela abordagem analítica, segmentando o sinal de fala em componentes menores, como fonemas ou difones, trabalhando assim com porções menores que tendem a ser processadas de forma mais rápida. A capacidade de reconhecer corretamente os fonemas permitiria construir uma interface que reconhecesse uma gama muito maior de palavras, abrindo mais o leque de possíveis aplicações para o reconhecimento de fala.

Há ainda a necessidade de explorar novas formas de pré-processamento, visando melhorar a qualidade do vetor característico do padrão. As formas empregadas, médias e FFT, mostraram-se frágeis a algumas variações comuns no sinal de fala, como entonação e timbre, indicando que há neste ponto uma possibilidade de melhoras significativas. Empregar a FFT sobre outras condições de análise da fala (como fonemas) ou utilizar-se da transformada *wavelet*[COD92], são formas de pré-processar que merecem atenção em trabalhos futuros.

6. BIBLIOGRAFIA

[ALL90] ALLEN,J. Speech recognition. In **Encyclopedia of Artificial Intelligence**, Vol 2, p. 1065-9. Editor-chefe Stuart C. Shapiro. New York: Wiley, 1990.

[ANA92] Analistas afirmam que o Windows dominará o mercado. In **BYTE** Brasil, vol 1, nr. 12, p. 18-27, Dez 1992.

[ANI93] ANICK, Peter G. Integrating natural language processing and information retrieval in a troubleshooting help desk. In **IEEE Expert**, p. 9-17, Dec 1993.

[BAR88] BARTHET, Marie-France. **Logiciels interactifs et ergonomie: modèles e méthodes de conception.** Paris: Dunod, 1988.

[BEL92] BELLONE, Roger. L'Ordinateur vocal interroge et répond. In **Science & Vie**, n° 894, p. 152. Paris: Excelsior , Mars 1992.

[BEZ92] BEZDEK, James C. , PAL, Sankar K. **Fuzzy models for pattern recognition: methods that search for structures in data.** New York: IEEE Press, 1992.

[BLU92] BLUM, A. **Neural networks in C++.** New York: John Wiley & Sons, 1992.

[BRI90] BRISCOE,E.J. Speech understanding . **In Encyclopedia of Artificial Intelligence.** Editor-chefe Stuart C. Shapiro. Vol 2, p. 1076-82. New York: Wiley, 1990.

[BRU95] BRUNS, Fábio Augusto. **Protótipo para o reconhecimento de palavras através da fala.** Trabalho de Conclusão de Curso em Ciências da Computação - Bacharelado. Professor-orientador: Marcel Hugo. Blumenau: FURB, Jul 1995.

[CAR88] CARPENTER, Gail A., GROSSBERG, Stephen. The ART of adaptative pattern recognition by a self-organizing neural network. In **IEEE Computer**, p. 77-88, Mar 1988.

[CAR92] CARRIJO, Gilberto Arantes, FIGUEIREDO, Maria Grace Silva. Reconhecimento de palavra isolada utilizando quantização de vetores e redes neurais artificiais. In **Ciência e Engenharia**, Ano 1, Nro. 2, p. 89-99. Uberlândia: Universidade Federal de Uberlândia - Centro de Ciências Exatas e Tecnologia, Jul-Dez/1992.

[CAU92] CAUDILL, Maureen. Kinder, gentler computing. In **BYTE**, p.135-50, Apr 1992.

[COD92] CODY, Mac A. The fast wavelet transform. In **Dr. Dobb's Journal**, p. 16-28, Apr 1992.

[COU90] COUTAZ, Joelle. **Interfaces homme-ordinateur: conception et réalisation**. Paris: Dunod, 1990.

[CRA93] CRANE, Hewitt D., RTISCHEV, Dimitry. Pen and voice unite. In **BYTE**, p. 98-102, Oct 1993.

[DAS92] DAS, Subrata , NADAS, Arthur. The power of speech . In **BYTE**, p.151-60, Apr 1992.

[FIS87] FISCHLER, Martin A., FIRSCHEIN, Oscar. **Intelligence: the eye, the brain, and the computer**. Menlo Park: Addison-Wesley, 1987.

[FRE81] FREIRE, Paulo. **Ação cultural para a liberdade** . 5ª Ed. Rio de janeiro: Paz e Terra, 1981.

[FRI94] FRITZ, Mark. Speech recognition finally gets some respect. In **Electronic Business Buyer**, Vol 20, nb 6, p.28, June 1994.

[FU82] FU, K.S. **Syntactic pattern recognition and applications**. Englewood Cliffs: Prentice Hall, 1982.

[FU83] FU, K.S. Pictorial pattern recognition for industrial inspection. In **Pictorial Data Analysis**, p. 335-49. Editado por R. M. Haralick. Berlin: Springer-Verlag, 1983.

[HUG92] HUGO, Marcel. **Construção de um protótipo de software comandado por voz**. Trabalho de Conclusão de Curso em Ciências da Computação - Bacharelado. Professor-orientador: Paulo de Tarso Mendes Luna. Blumenau: FURB, Nov 1992.

[HUN92] HUNTSBERGER, Terrance L., PONGSAK, Ajjimarangsee. Parallel self-organizing feature maps for unsupervised pattern recognition. In BEZDEK, James C. , PAL, Sankar K. **Fuzzy models for pattern recognition: methods that search for structures in data**, p. 483-95. New York: IEEE Press, 1992.

[INS92] INSTITUTE OF INDUSTRIAL ENGINEERS. America's cup contender's on-board system integrates communications. In **Industrial Engineering**, p.19, Atlanta-USA, Aug 1992.

[IYE91] IYENGAR, S.S., KASHYAP, R.L. Neural networks: a computational perspective. In ANTOGNETTI, Paolo, MILUTINOVIC, Veljko. **Neural networks: concepts, applications, and implementations**. Vol II, p. 1-30. Englewood Cliffs: Prentice Hall, 1991.

[JEN91] JENSON, Tory. From telephone to database . In **Datamation**, p. 46-8, Dec/15/1991.

[KAN90] KANAL, L.N., DATTATREYA, G.R. Pattern recognition . In **Encyclopedia of Artificial Intelligence**, Vol 2, p. 720-

9. Editor-chefe Stuart C. Shapiro. New York: Wiley, 1990.

[KOH88] KOHONEN, Teuvo. An introduction to neural computing. In **Neural Networks**, Vol.1 , p. 3-16. USA: Pergamon Journals, 1988.

[KOH90] KOHONEN, Teuvo. The self-organizing map. In **Proceedings of the IEEE**, vol 78, nb 9, p. 1464-80, Sep 1990.

[LAP88] LAPOLLI, Édis Mafra. **Escolha de rotas em Centrais de Informação de Fretes.** Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção. Professor-orientador: Ricardo Miranda Barcia. Florianópolis: UFSC, 1988.

[LAW92] LAWRENCE, Janet. **Introduction to neural networks and expert systems.** 4ª ed, Jan 1992.

[LIP87] LIPPMANN, Richard P. An introduction to computing with neural nets. In **IEEE ASSP Magazine**, vol 3, nb 4, p. 4-22, Apr 1987.

[MAI85] MAIA, Eleonora Motta. **No reino da fala: a linguagem e seus sons.** Ática, 1985.

[MAR90] MAREN, Alliana, HARSTON, Craig, PAP, Robert. **Handbook of neural computing applications.** San Diego: Academic Press, 1990.

[MAT91] MATRAS, Jean-Jacques. **O som.** São Paulo: Martins Fontes, 1991.

[MEI93] MEISEL, William S. Talk to your computer. In **BYTE**, October 1993, p.113-20.

[MEZ93] MEZICK, Dan. Pen computing catches on. In **BYTE**, Oct 1993, p. 105-12.

[MIC94] MICROSOFT CORPORATION. **Windows 3.1 SDK: multimedia programmer's guide.** Microsoft Development Library, Disk 7, CD-ROM. Apr 1994.

[MUN94] MUNNIK, Josha and OOSTENDORP, Eric. **The Sound Blaster book.** Alameda: SYBEX, 1994.

[MUS93] MUSSER, John. A multimedia class library for Windows. In **Dr. Dobbb's Journal**, p. 84-90, Jul 1993.

[PEZ93] PEZZI, Silvana. **Uma interface para o processo de informatização das Centrais de Informações de Fretes.** Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção. Professor-orientador: Ricardo Miranda Barcia. Florianópolis: UFSC, 1993.

[PIE87] PIERREL, Jean-Marie. **Dialogue oral homme-machine**. Paris: Hermes, 1987.

[QUA94] QUAIN, John R. Windows Sound System 2.0 targets business users. In **PC Magazine**, Feb 1994, p.52.

[RIC88] RICH, Elaine. **Inteligência artificial**. Tradução: Newton Vasconcellos. São Paulo: McGraw-Hill, 1988.

[RIC93] RICH, Elaine, KNIGHT, Kevin. **Inteligência artificial**. Tradução: Maria Cláudia Santos Ribeiro Ratto. São Paulo: Makron Books, 1993.

[RUD93] RUDNICKY, Alex. Matching the input mode to the task. In **BYTE**, Oct 1993. p. 100.

[SAV88] SAVADOVSKY, Pedro. **Introdução ao projeto de interfaces em linguagem natural**. São Paulo: SID Informática, 1988.

[SCA86] SCAPIN, Dominique L. **Guide ergonomique de conception des interfaces homme-machine**. Institut National de Recherche en Informatique et en Automatique. Rapports Techniques 77. France, Oct 1986.

[SEY94] SEYBOLD PUBLICATIONS. Speech recognition: state of the art improves. In **The Seybold Report on Desktop Publishing**, v8, n7, p.1, March 07 1994.

[SLO80] SLOBIN, Dan Isaac. **Psicolingüística** . São Paulo: Nacional, Universidade de São Paulo, 1980.

[STO93] STOLZ, Alex. **The Sound Blaster book**. USA: Abacus, 1993.

[THO81] THOMAS, J.C., CARROLL, J.M. Human factors in communication. In **IBM System Journal**, vol 20, nb 2, 1981, p. 237-67.

[TEB95] TEBBUTT, David. In touch with tomorrow. In **PC PRO**, Feb 1995, p. 206-11. Dennis Publication, United Kingdom.

[THR91] THRO, Ellen. **The artificial intelligence dictionary**. The Lance. A. Leventhal Microtrend Series. San Marcos: Microtrend, 1991.

[TUC92] TUCKER, R. **Voice activity detection using a periodicity measure**. In IEEE Proceedings, vol 139, n° 4, Aug 1992.

[VER92] VERHAEGHE, Bart. Toward continous-speech recognition. In **BYTE**, Apr 1992, p. 158.

[WAN93] WANG, DeLiang. Pattern recognition: neural networks in perspective. In **IEEE Expert**, Aug 1993, p. 52-60.

MARCAS REGISTRADAS

Audio Info Engine: Gralin Associates Inc.
CI-500: SIL - Summer Institute of Linguistics.
Computerfone III: Suncoast Systems Inc.
DragonDictate: Dragon Systems.
Fusion CD 16: Media Vision.
MS-DOS, Microsoft Visual Basic, Windows, Windows NT, Windows Sound System: Microsoft Co.
Personal Dictation System, OS/2 e Tangora: IBM Co.
Sound Blaster: Creative Labs.
Speech Master System: Speech Soft Inc.
Verbex Voice Systems: Edison Co.
Voice Information Processing Server: Octel Communication Co.
Vox'Scrib Base: DECICOM.

O autor declara estar usando as marcas citadas nesta obra apenas para fins acadêmicos, em benefício do dono da marca, sem a intenção de infringir as regras de seu uso.